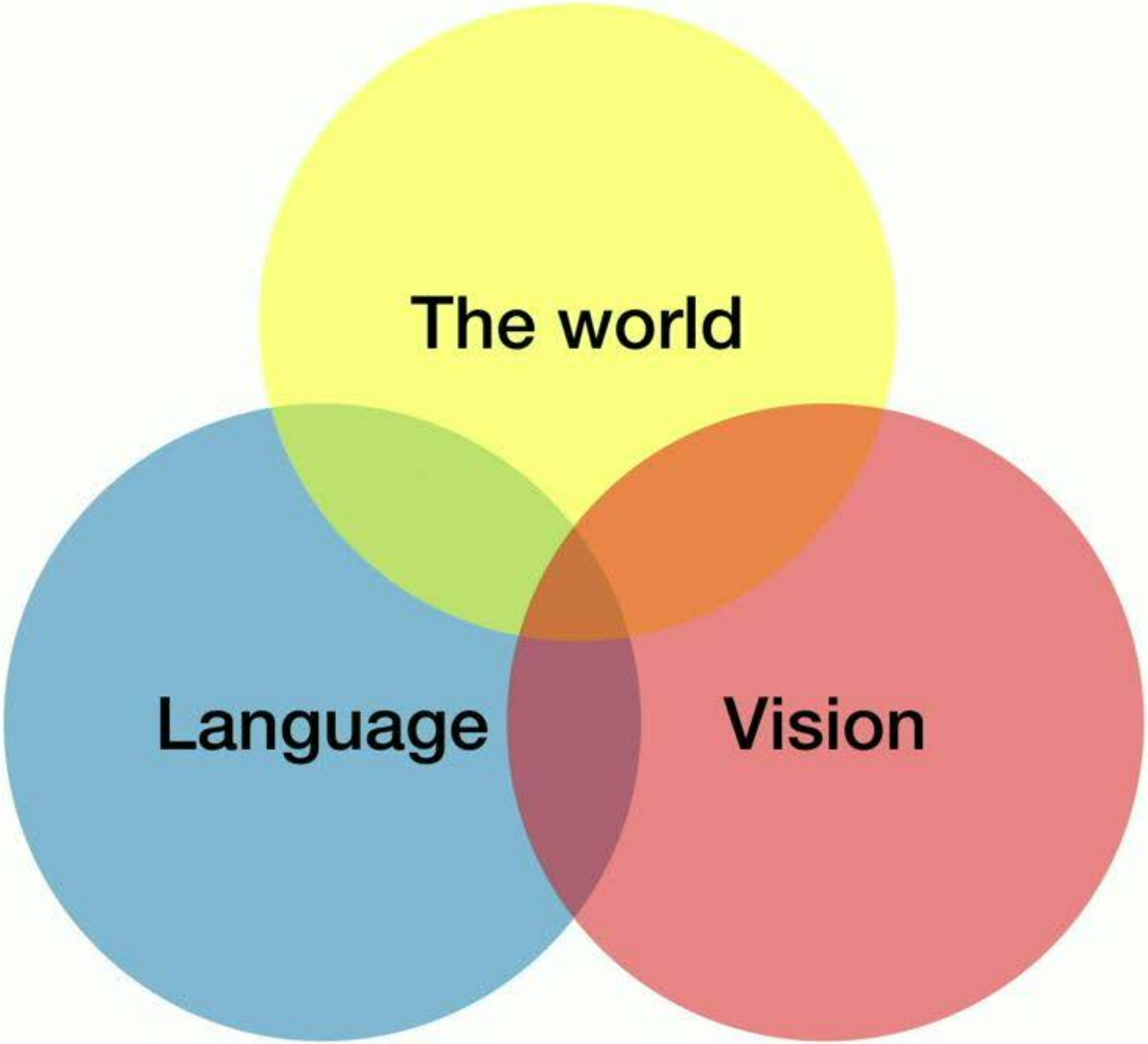
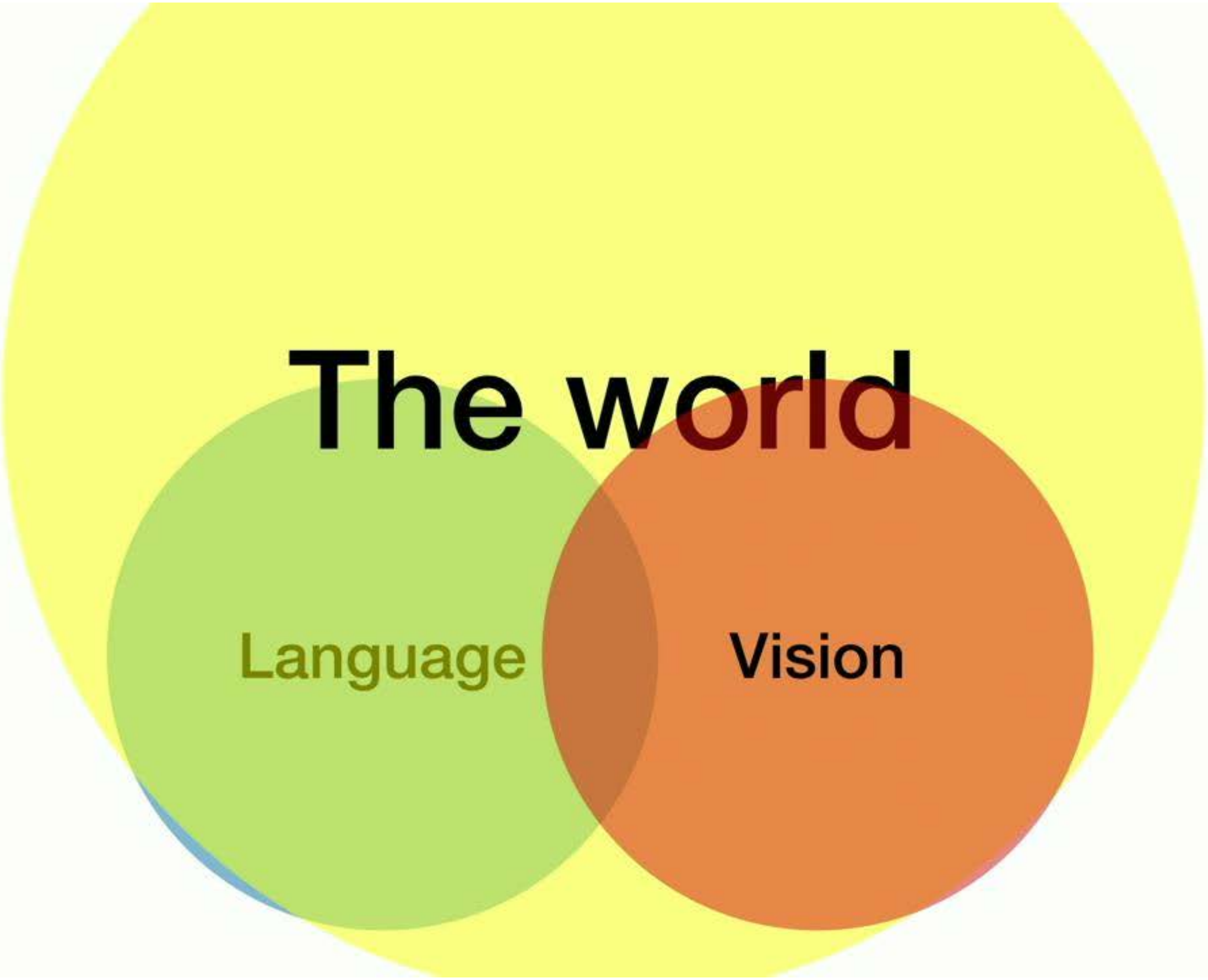
Building Robust Datasets for Commonsense Reasoning

Rowan Zellers

rowanzellers.com


The world
Language
Vision


The world
Language
Vision

A man has a few tools and is pumping his car up so he can take off the tire.
He…

A man has a few tools and is pumping his car up so he can take off the tire.
He…

A. stops on the front of the bike and moves it to the left.

B. gets out of the car while leaving the engine running.

C. uses the tool to take off all of the nuts one by one.

D. goes down from the cars, landing straight in.

A man has a few tools and is pumping his car up so he can take off the tire. He…

A. stops on the front of the bike and moves it to the left.

B. gets out of the car while leaving the engine running.

C. uses the tool to take off all of the nuts one by one.

D. goes down from the cars, landing straight in.

A man has a few tools and is pumping his car up so he can take off the tire.
He…

A. stops on the front of the bike and

This is natural language inference that requires commonsense reasoning!

one by one.

D. goes down from the cars, landing straight in.

SWAG: A Large-Scale Adversarial Dataset for Grounded Commonsense Inference (emnlp18)

Our contributions with SWAG

- SWAG: Natural Language Inference + Commonsense Reasoning

Our contributions with SWAG

- SWAG: Natural Language Inference + Commonsense Reasoning

- Adversarial Filtering

Scope of Natural Language Inference (NLI)

Much of today's NLI requires only linguistic knowledge, without as much commonsense reasoning.

Bowman et al., 2015, Wang et al., 2018

Scope of Natural Language Inference (NLI)

Much of today's NLI requires only linguistic knowledge, without as much commonsense reasoning. ***Examples from SNLI:***

A woman is talking on the phone while standing next to a dog.

?

A woman is talking on the phone.

Bowman et al., 2015, Wang et al., 2018

Scope of Natural Language Inference (NLI)

Much of today's NLI requires only linguistic knowledge, without as much commonsense reasoning. ***Examples from SNLI:***

A woman is talking on the phone while standing next to a dog.

?

A woman is talking on the phone.

Entailment

Intersectivity

Bowman et al., 2015,
Wang et al., 2018

Scope of Natural Language Inference (NLI)

Much of today's NLI requires only linguistic knowledge, without as much commonsense reasoning. ***Examples from SNLI:***

A woman is talking on the phone while standing next to a dog. ? → A woman is talking on the phone.

Entailment

Intersectivity

A man is doing tricks on a skateboard. ? → Nobody is doing tricks.

Bowman et al., 2015, Wang et al., 2018

Scope of Natural Language Inference (NLI)

Much of today's NLI requires only linguistic knowledge, without as much commonsense reasoning. ***Examples from SNLI:***

A woman is talking on the phone while standing next to a dog.

A woman is talking on the phone.

Entailment

Intersectivity

A man is doing tricks on a skateboard.

Nobody is doing tricks.

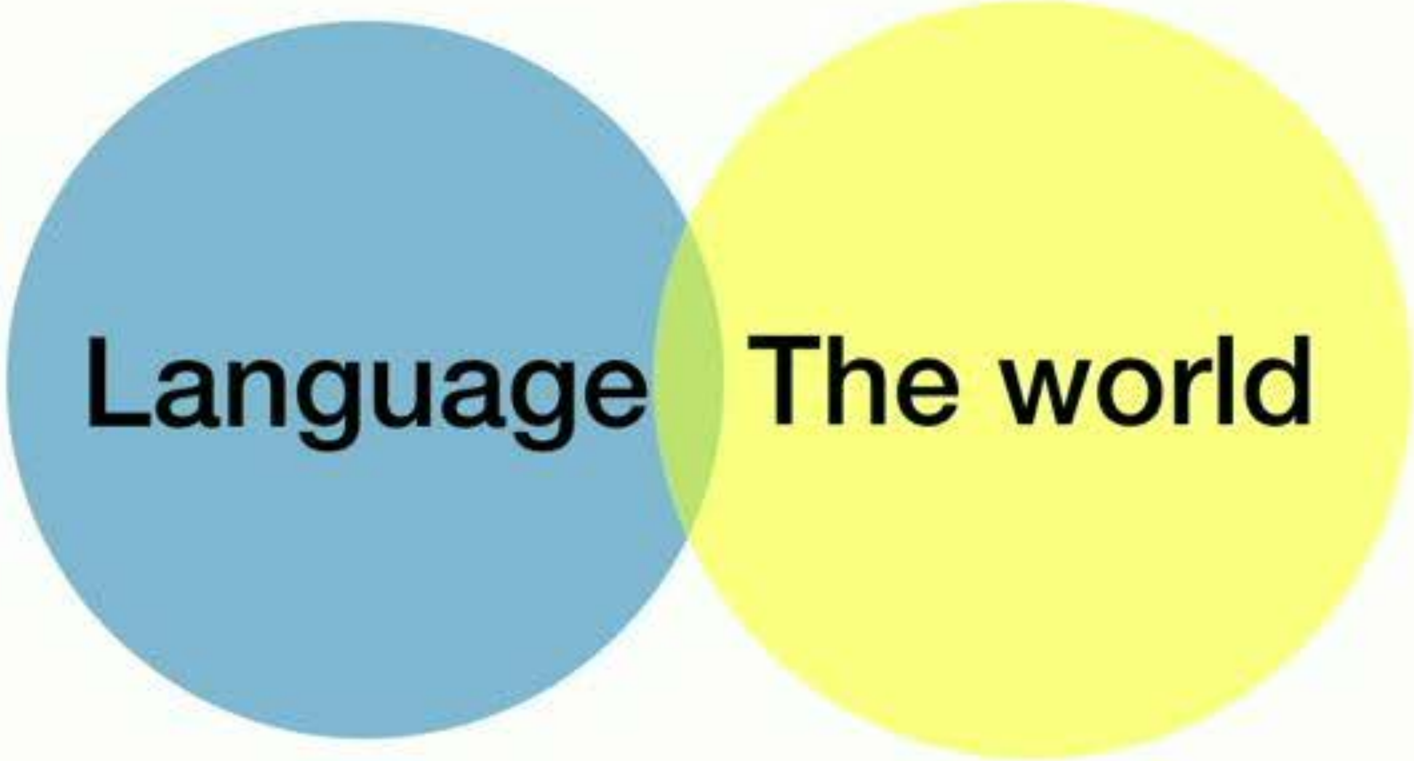
Contradiction

Quantifiers

Bowman et al., 2015, Wang et al., 2018

Re-emphasizing Commonsense in NLI

Natural Language Inference
was supposed to be:



(Dagan et. al, 2006; LoBue and Yates, 2011)

Situations with Adversarial Generations (SWAG)

A man has a few tools and is pumping his car up so he can take off the tire. He

t

ACTIVITYNET

LSMDC

Krishna et al., 2017, Rohrbach et al., 2016

Situations with Adversarial Generations (SWAG)

A man has a few tools and is pumping his car up so he can take off the tire. He

A. stops on the front of the bike and moves it to the left.
B. gets out of the car while leaving the engine running.
C. uses the tool to take off all of the nuts one by one.
D. goes down from the cars, landing straight in

t

t+1

ACTIVITYNET
LSMDC

Krishna et al., 2017, Rohrbach et al., 2016

Situations with Adversarial Generations (SWAG)

A man has a few tools and is pumping his car up so he can take off the tire. He

A. stops on the front of the bike and moves it to the left.
B. gets out of the car while leaving the engine running.
C. uses the tool to take off all of the nuts one by one.
D. goes down from the cars, landing straight in

SWAG is purely a natural language inference task: the video is never used.

ACTIVITYNET
LSMDC

Krishna et al., 2017, Rohrbach et al., 2016

Situations with Adversarial Generations (SWAG)

A man has a few tools and is pumping his car up so he can take off the tire. He

A. stops on the front of the bike and moves it to the left

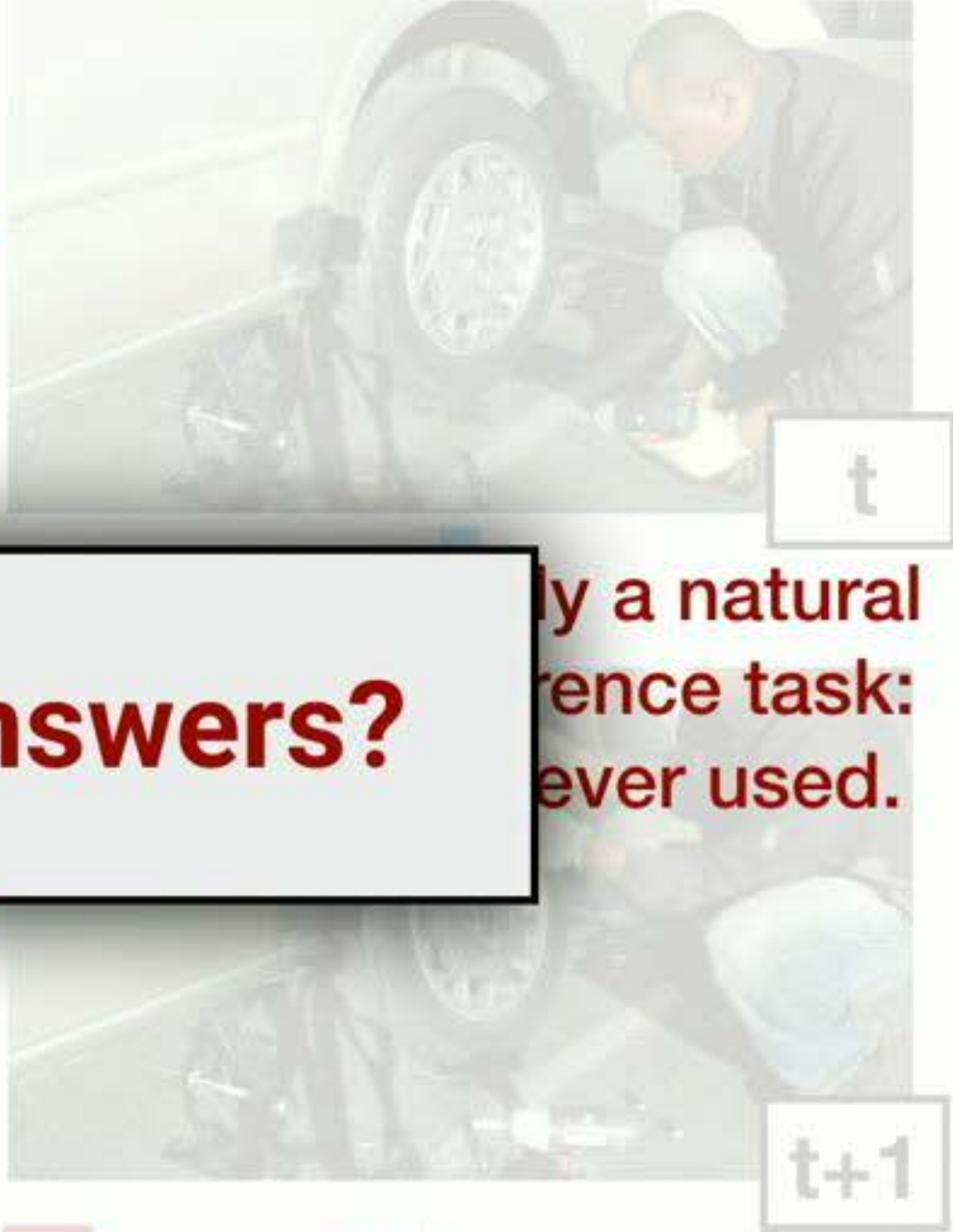
B. g... e...

C. u... nuts one by one.

D. goes down from the cars, landing straight in

How do we get the wrong answers?

...ly a natural ...rence task: ...ever used.



ACTIVITYNET
LSMDC

Krishna et al., 2017, Rohrbach et al., 2016

Our contributions

- SWAG: Natural Language Inference + Commonsense Reasoning

- Adversarial Filtering

Annotation Artifacts

Human written datasets are susceptible to *annotation artifacts*: stylistic patterns that give unwanted clues for the labels.

Schwartz et al., 2017
Gururangan et al., 2018
Poliak et al., 2018

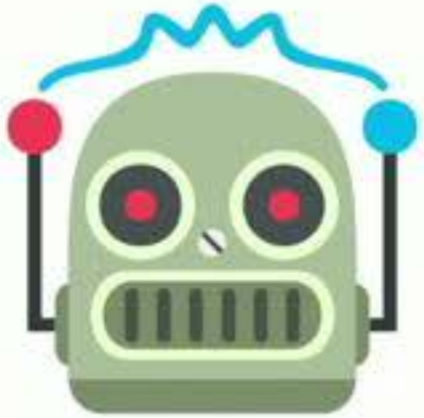
Annotation Artifacts

Human written datasets are susceptible to *annotation artifacts*: stylistic patterns that give unwanted clues for the labels.

Schwartz et al., 2017
Gururangan et al., 2018
Poliak et al., 2018

A man is doing tricks on a skateboard. → ? → Nobody is doing tricks.

Contradiction



Annotation Artifacts

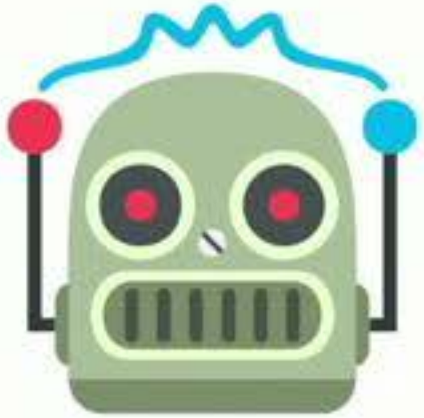
Human written datasets are susceptible to *annotation artifacts*: stylistic patterns that give unwanted clues for the labels.

Schwartz et al., 2017
Gururangan et al., 2018
Poliak et al., 2018

A man is doing tricks on a skateboard. → ? → Nobody is doing tricks.

Contradiction

For *Contradiction*, "Nobody" works…



Annotation Artifacts

Human written datasets are susceptible to *annotation artifacts*: stylistic patterns that give unwanted clues for the labels.

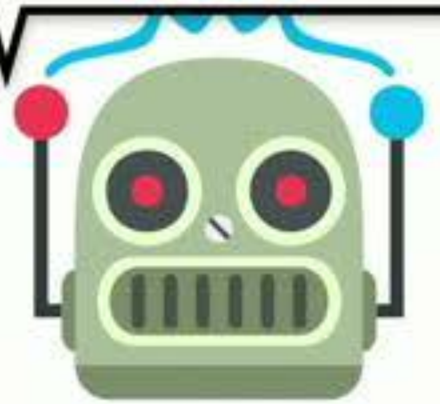
Schwartz et al., 2017
Gururangan et al., 2018
Poliak et al., 2018

A man is doing tricks on a skateboard. → ? → **Nobody** is doing tricks.

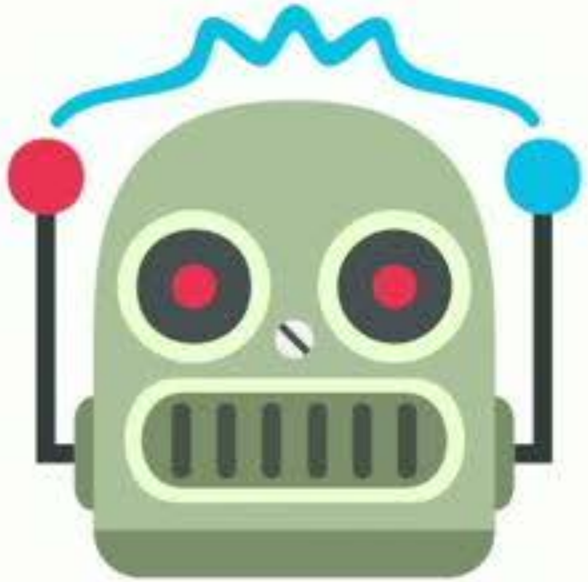
Contradiction

For *Contradiction*, "Nobody" works…

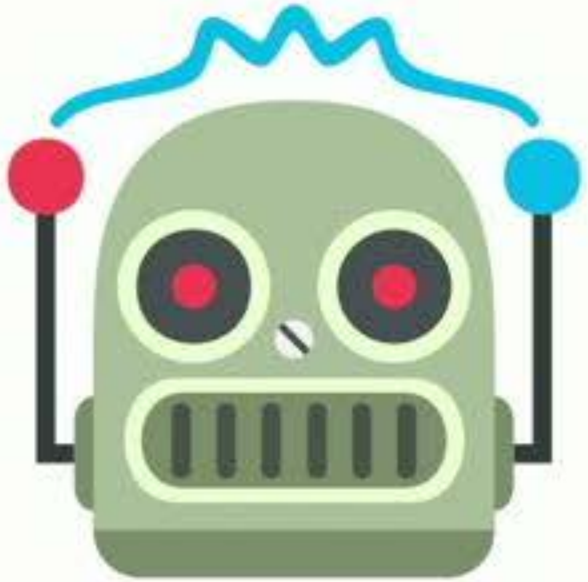
```
Nobody ->
Negation
```



A different approach…

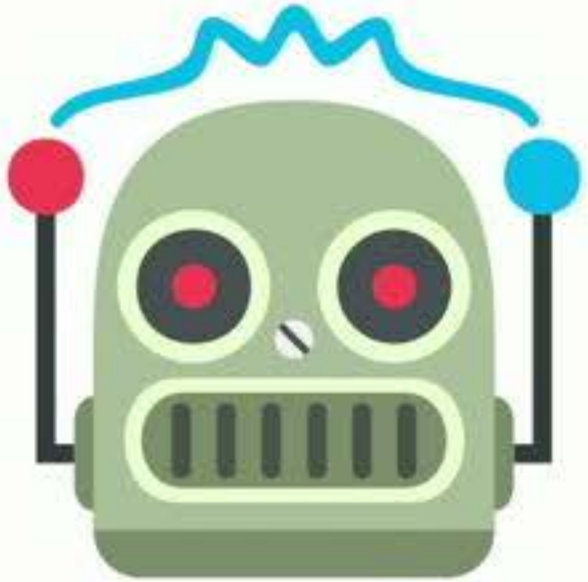


A different approach…



Let's have machines write the wrong endings!

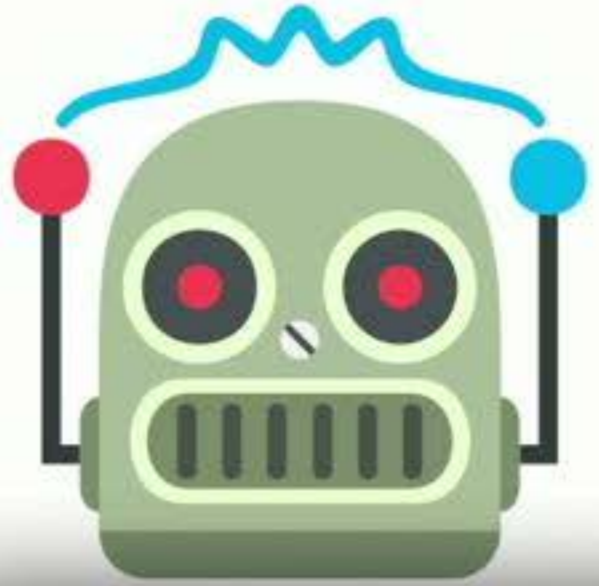
A different approach…



Let's have machines write the wrong endings!

Humans will **verify** that the original endings are better than the wrong ones.

A different approach…

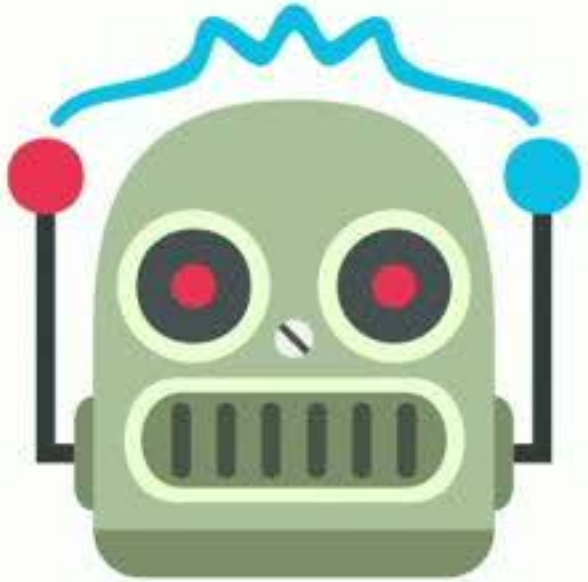


Let's have machines write the wrong endings!

Machine written text still has artifacts!

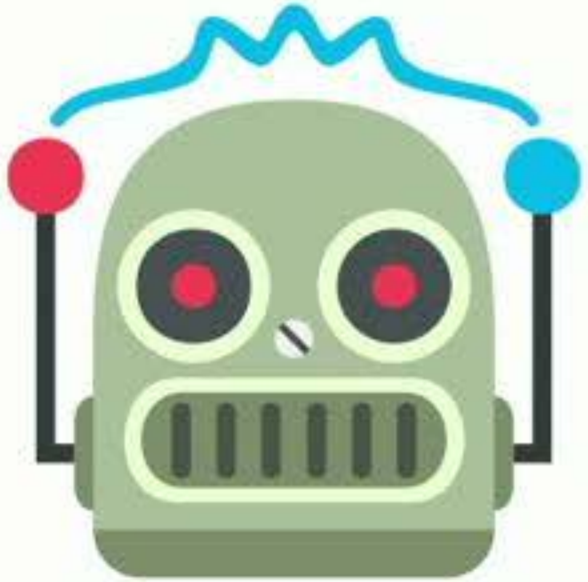
Humans will **verify** that the original endings are better than the wrong ones.

Adversarial Filtering

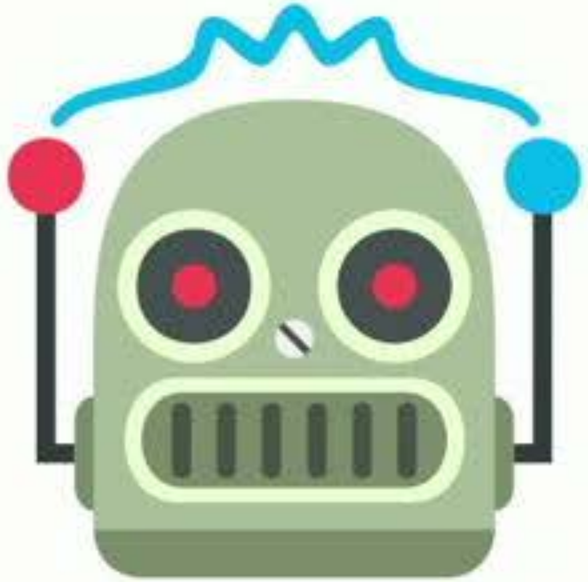


We'll *massively oversample* candidate endings.

Adversarial Filtering

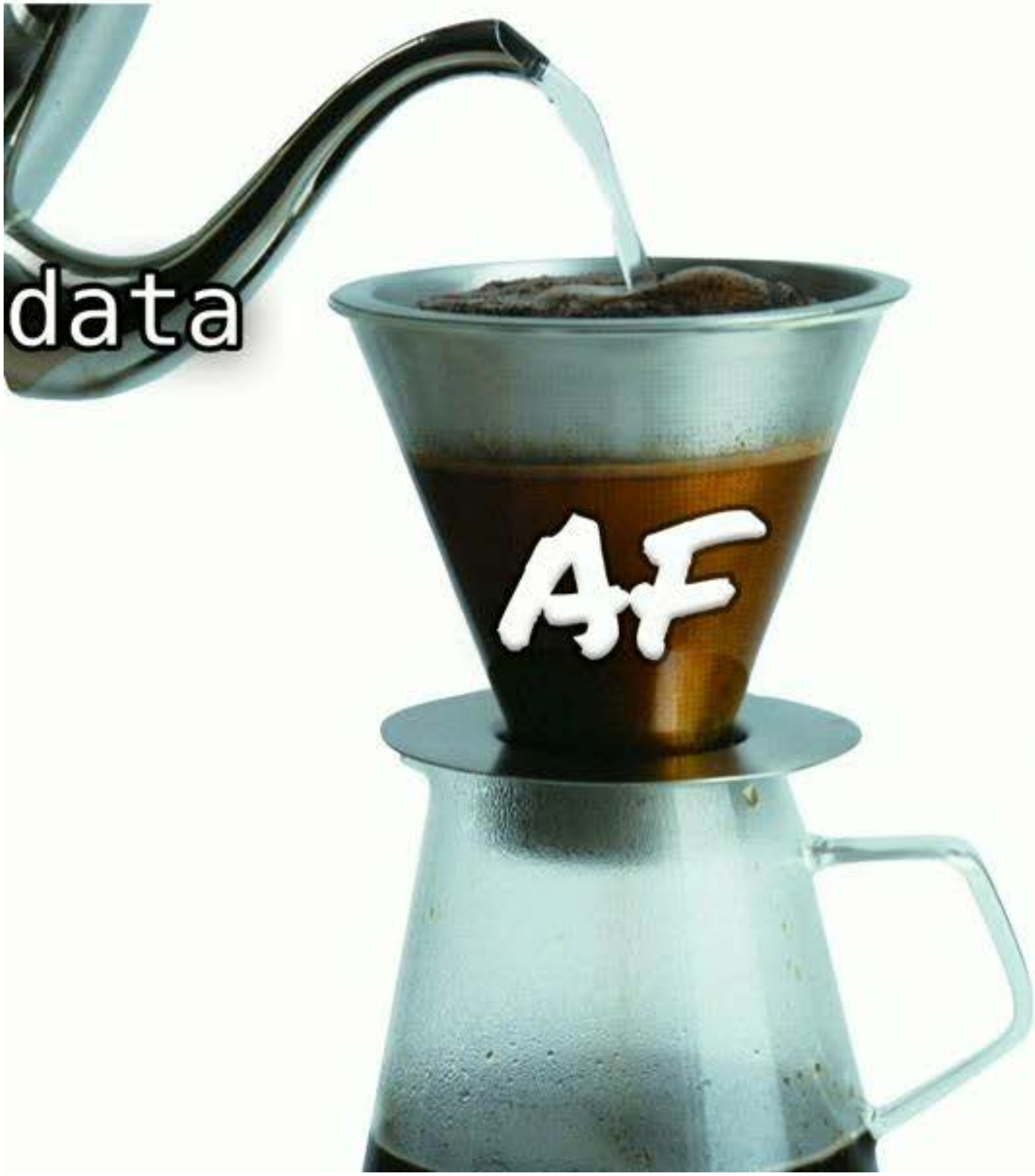


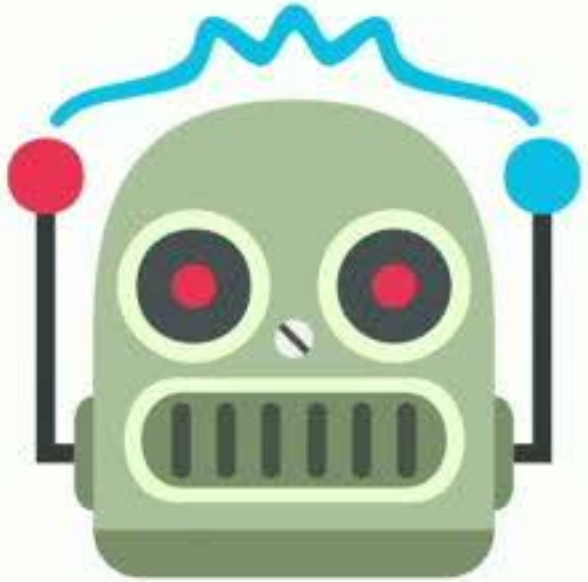
We'll *massively oversample* candidate endings.



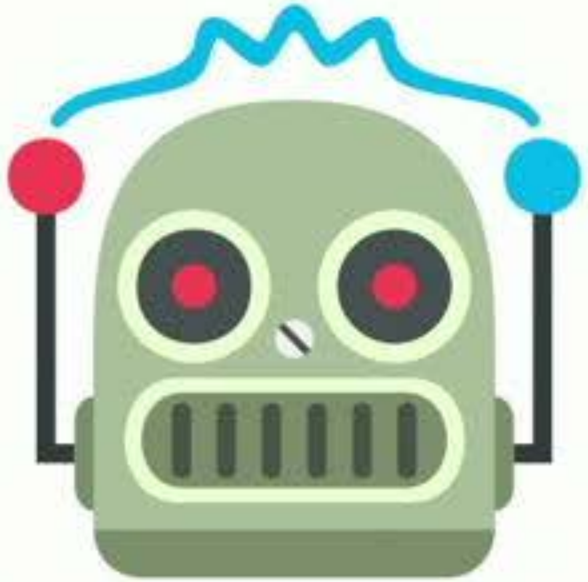
And we'll use more models to remove obvious artifacts!

Adversarial Filtering





We'll *massively oversample* candidate endings.



And we'll use more models to remove obvious artifacts!

Adversarial Filtering

A man has a few tools and is pumping his car up so he can take off the tire.

He uses the tool to take off all of the nuts one by one.

NP ***Ground truth VP***

Adversarial Filtering

A man has a few tools and is pumping his car up so he can take off the tire.

He uses the tool to take off all of the nuts one by one.

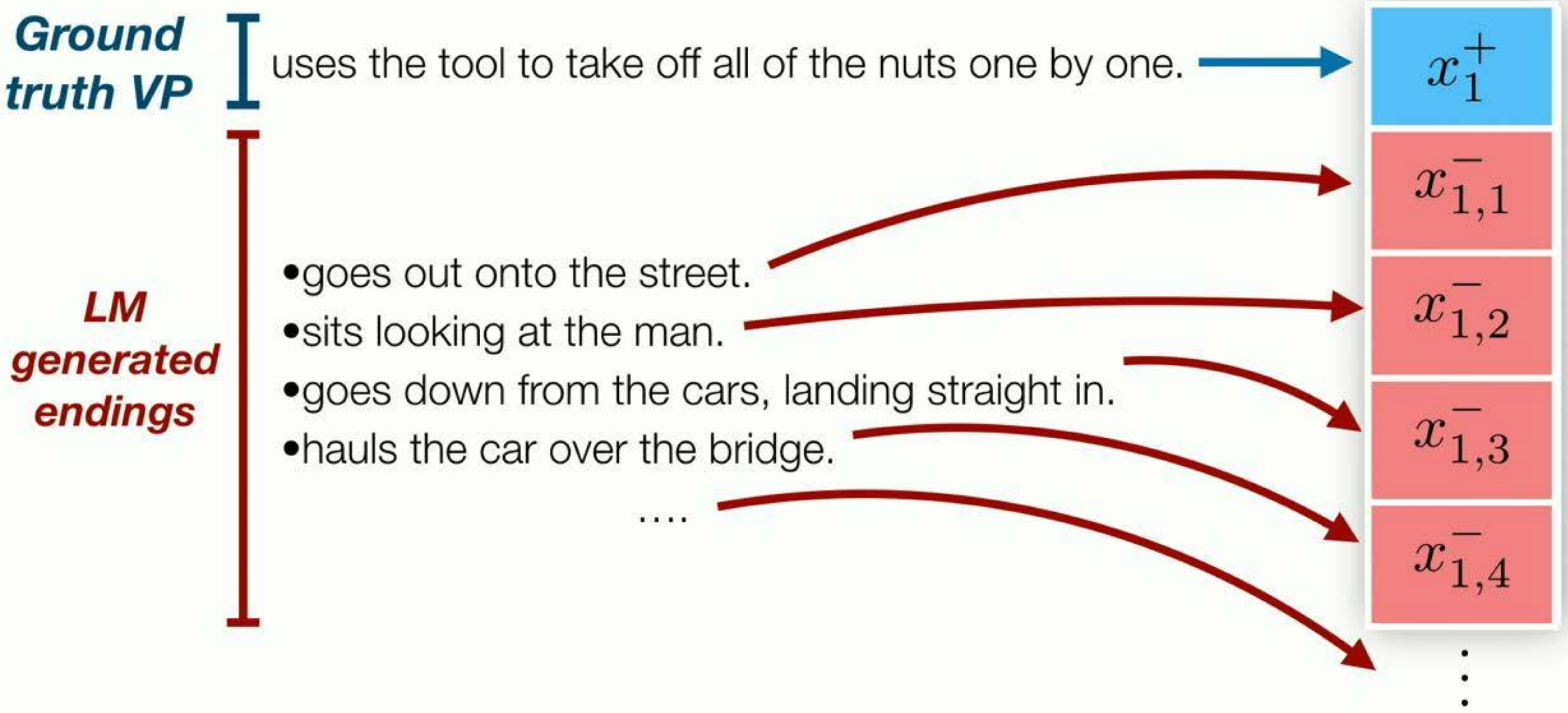
NP / **Ground truth VP**

LM generated endings

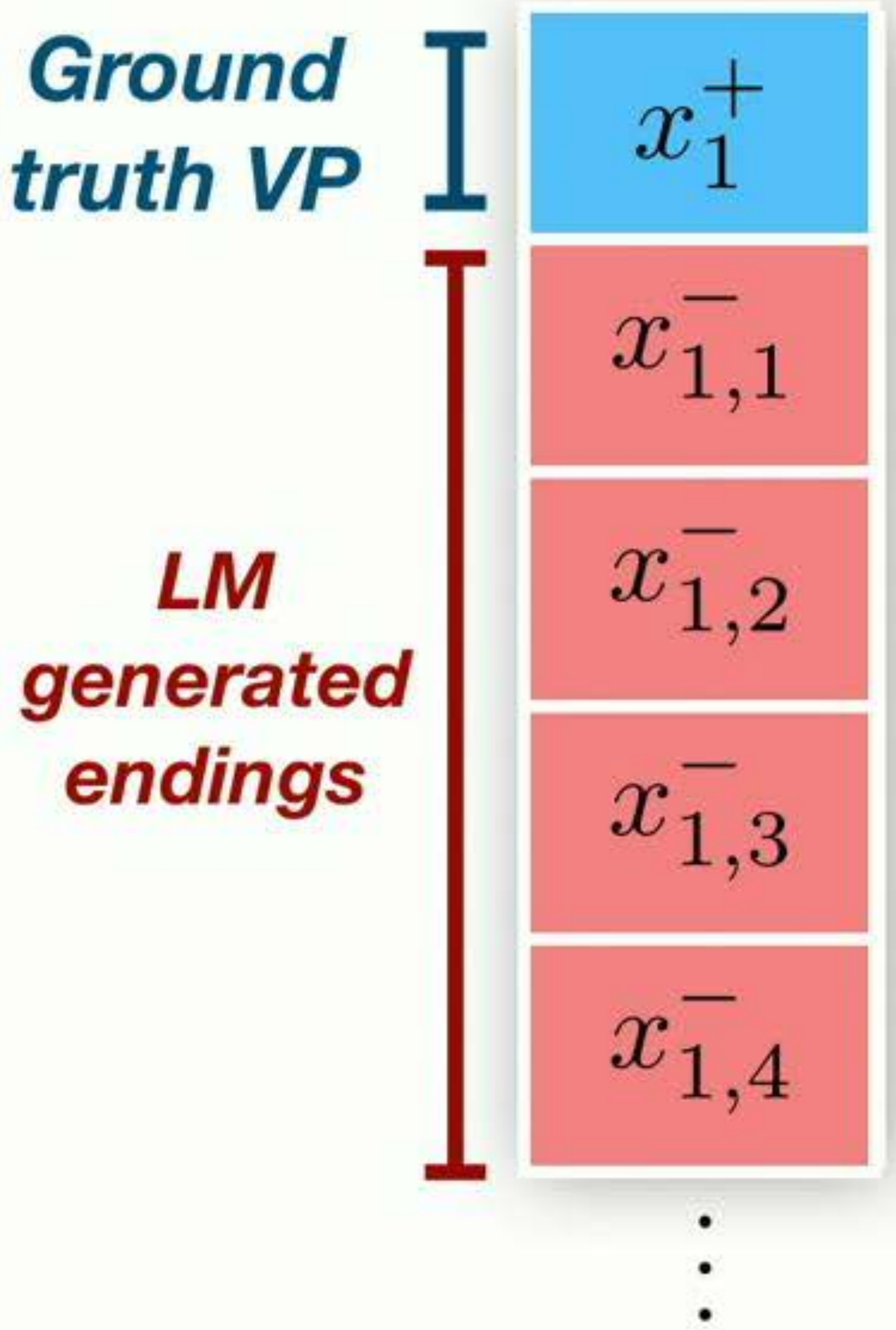
- goes out onto the street.
- sits looking at the man.
- goes down from the cars, landing straight in.
- hauls the car over the bridge.

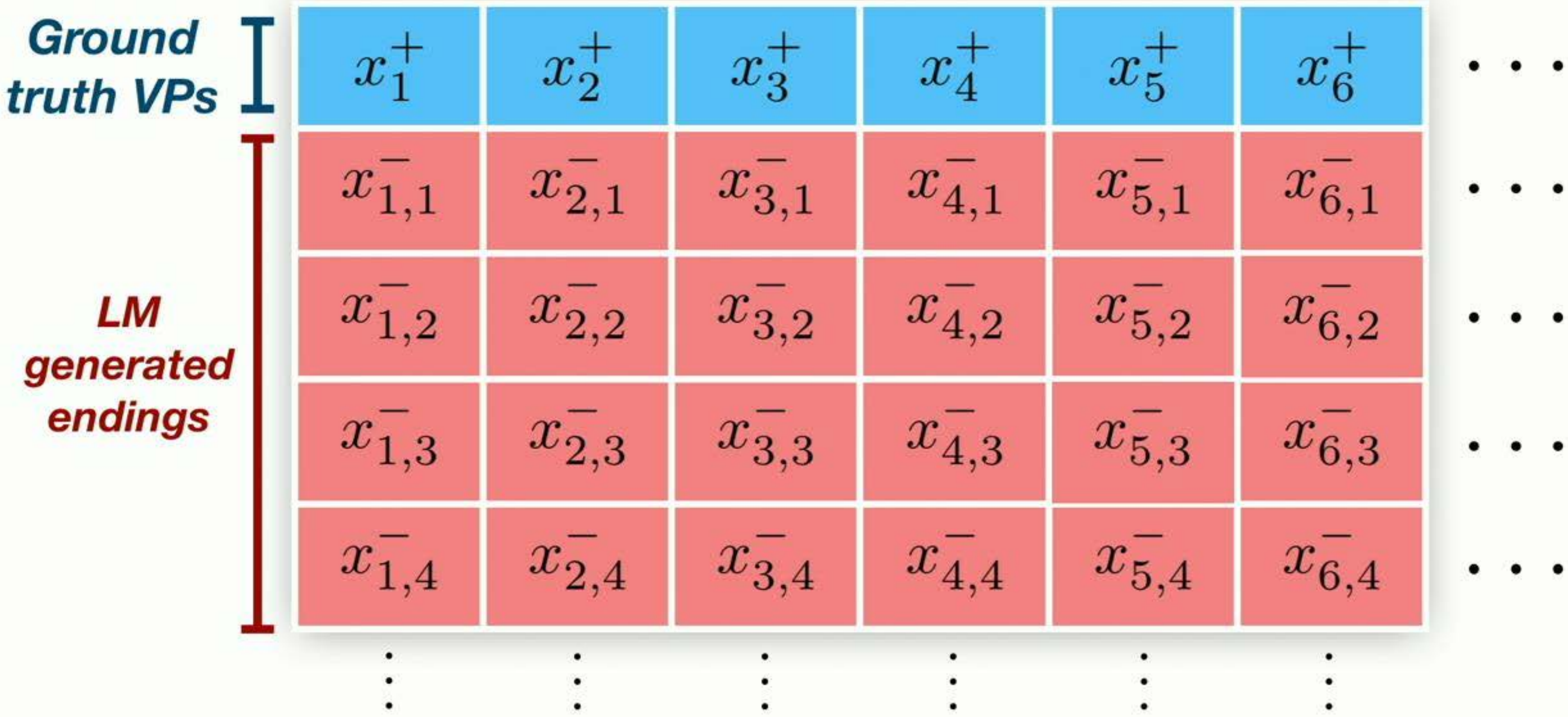
....

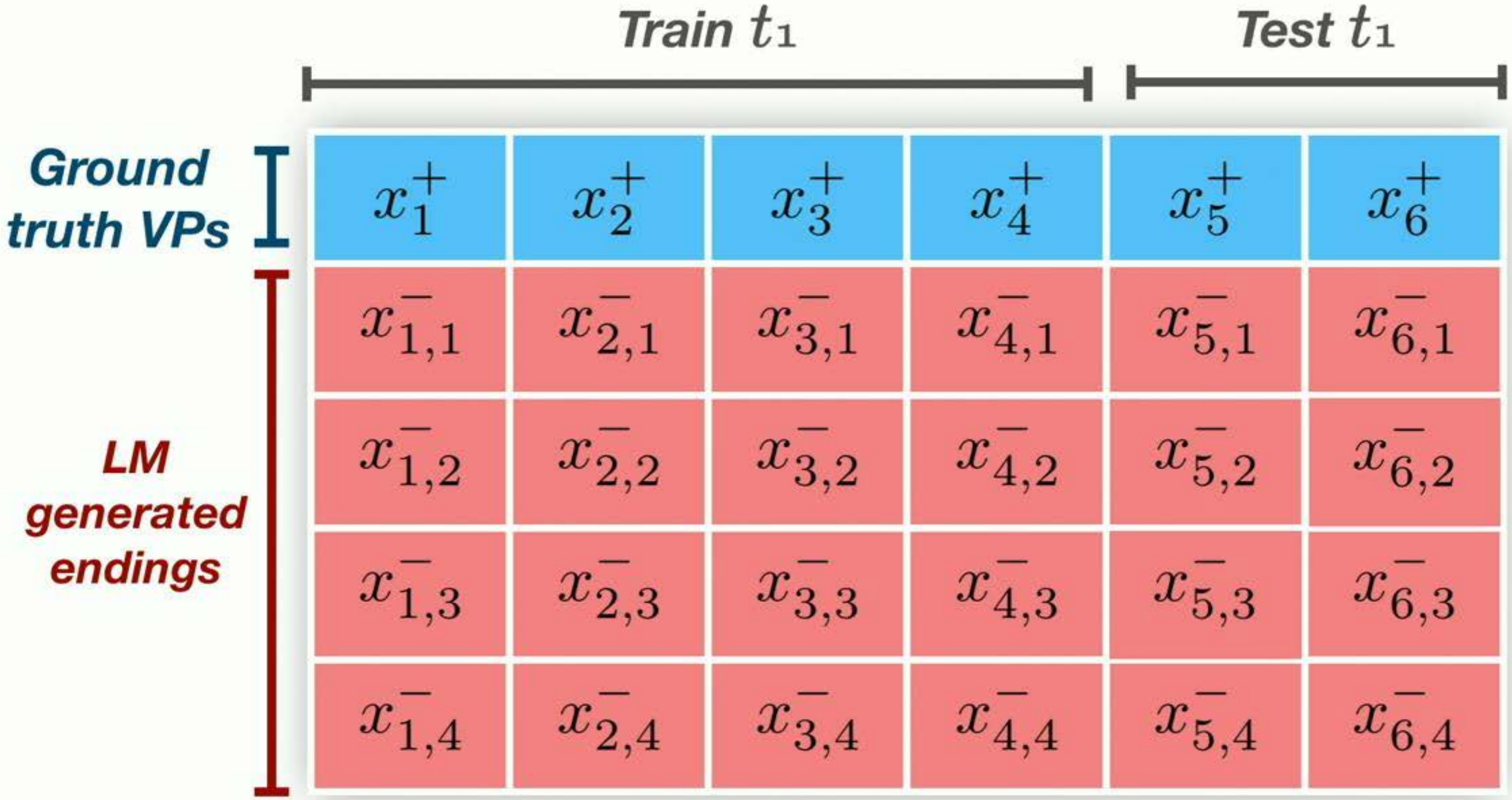
Adversarial Filtering

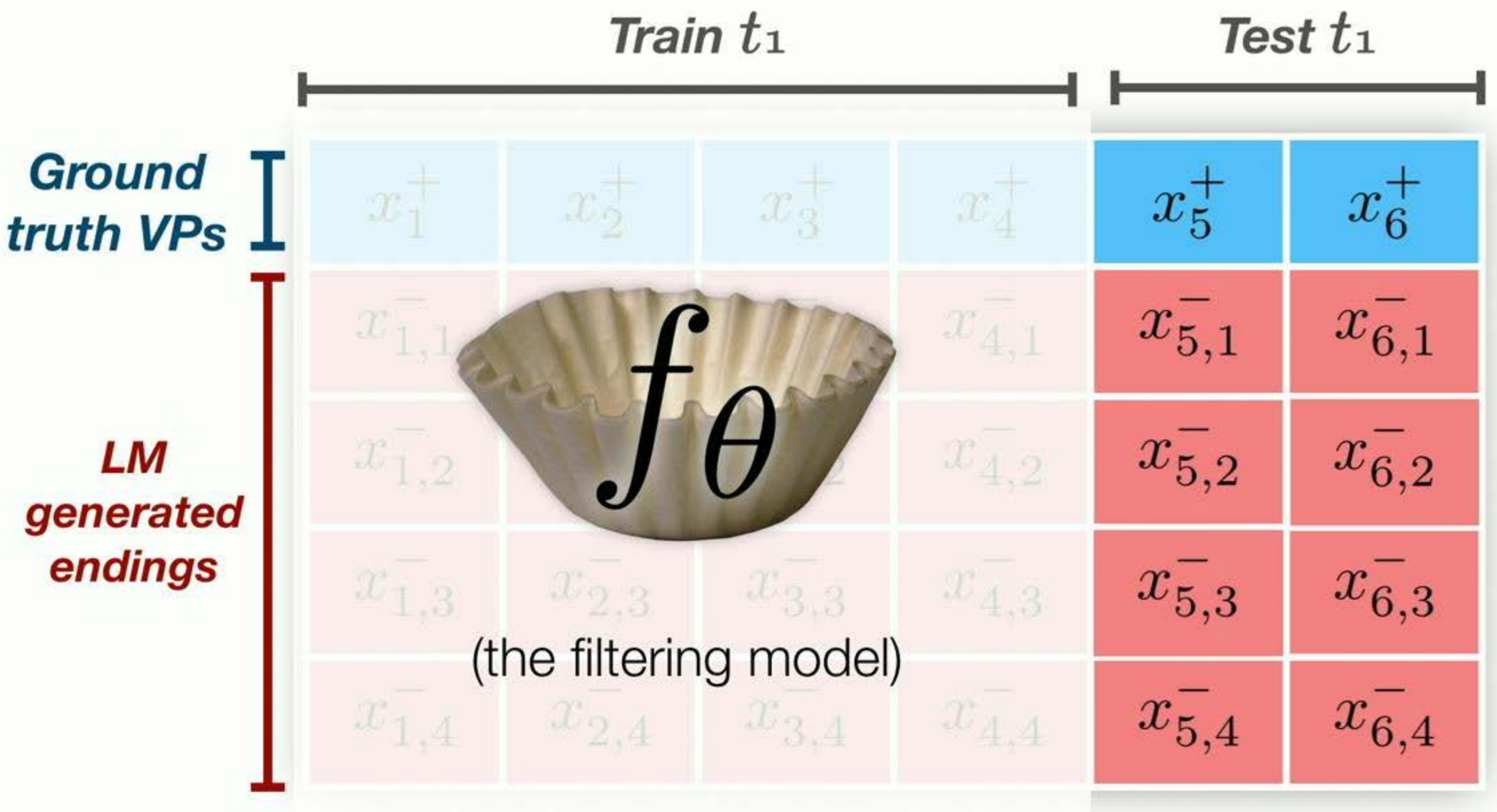


Adversarial Filtering

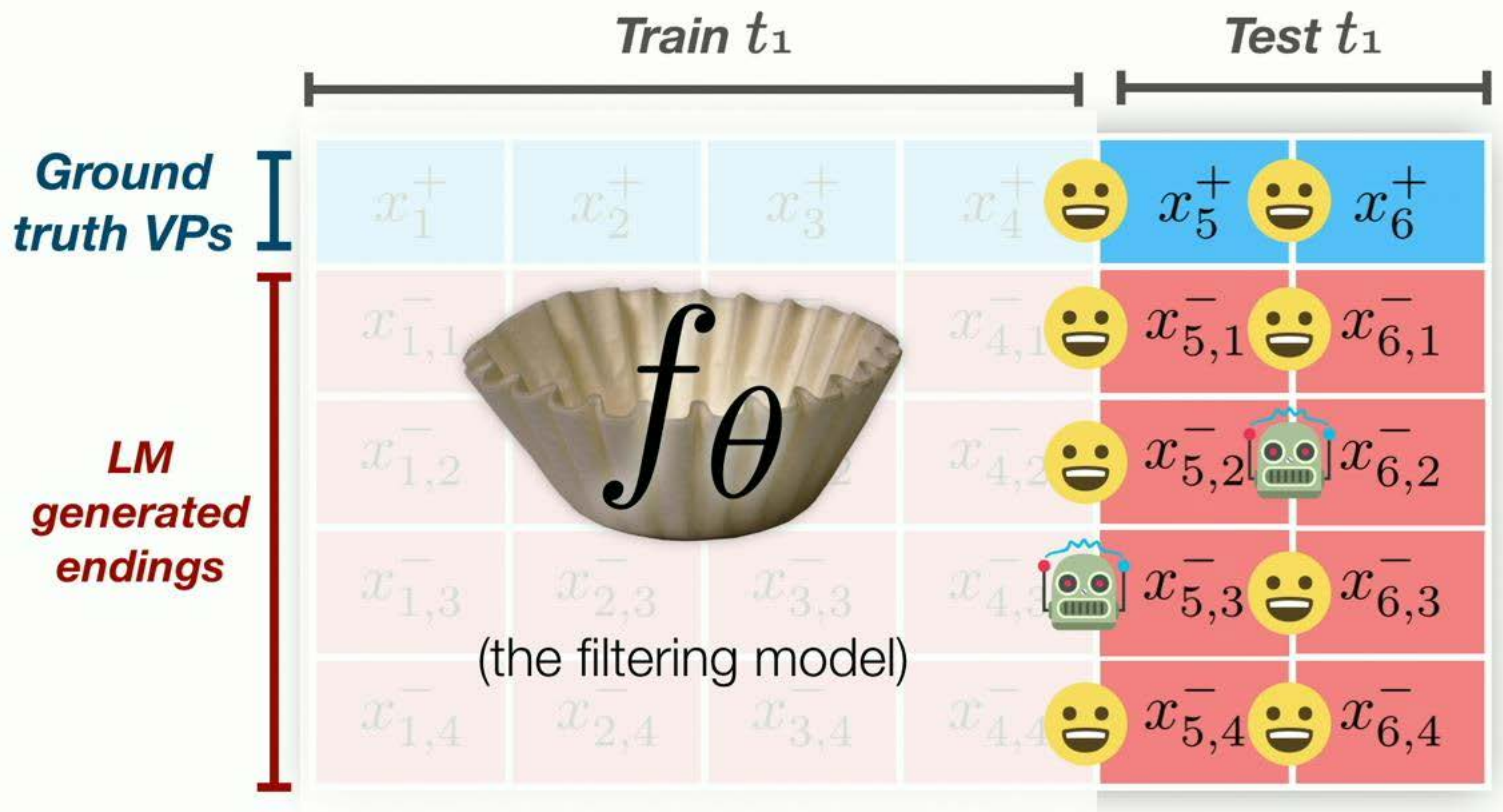



Ground truth VPs
x_1^+ x_2^+ x_3^+ x_4^+ x_5^+ x_6^+ ...
LM generated endings
$x_{1,1}^-$ $x_{2,1}^-$ $x_{3,1}^-$ $x_{4,1}^-$ $x_{5,1}^-$ $x_{6,1}^-$ ...
$x_{1,2}^-$ $x_{2,2}^-$ $x_{3,2}^-$ $x_{4,2}^-$ $x_{5,2}^-$ $x_{6,2}^-$ ...
$x_{1,3}^-$ $x_{2,3}^-$ $x_{3,3}^-$ $x_{4,3}^-$ $x_{5,3}^-$ $x_{6,3}^-$ ...
$x_{1,4}^-$ $x_{2,4}^-$ $x_{3,4}^-$ $x_{4,4}^-$ $x_{5,4}^-$ $x_{6,4}^-$ ...

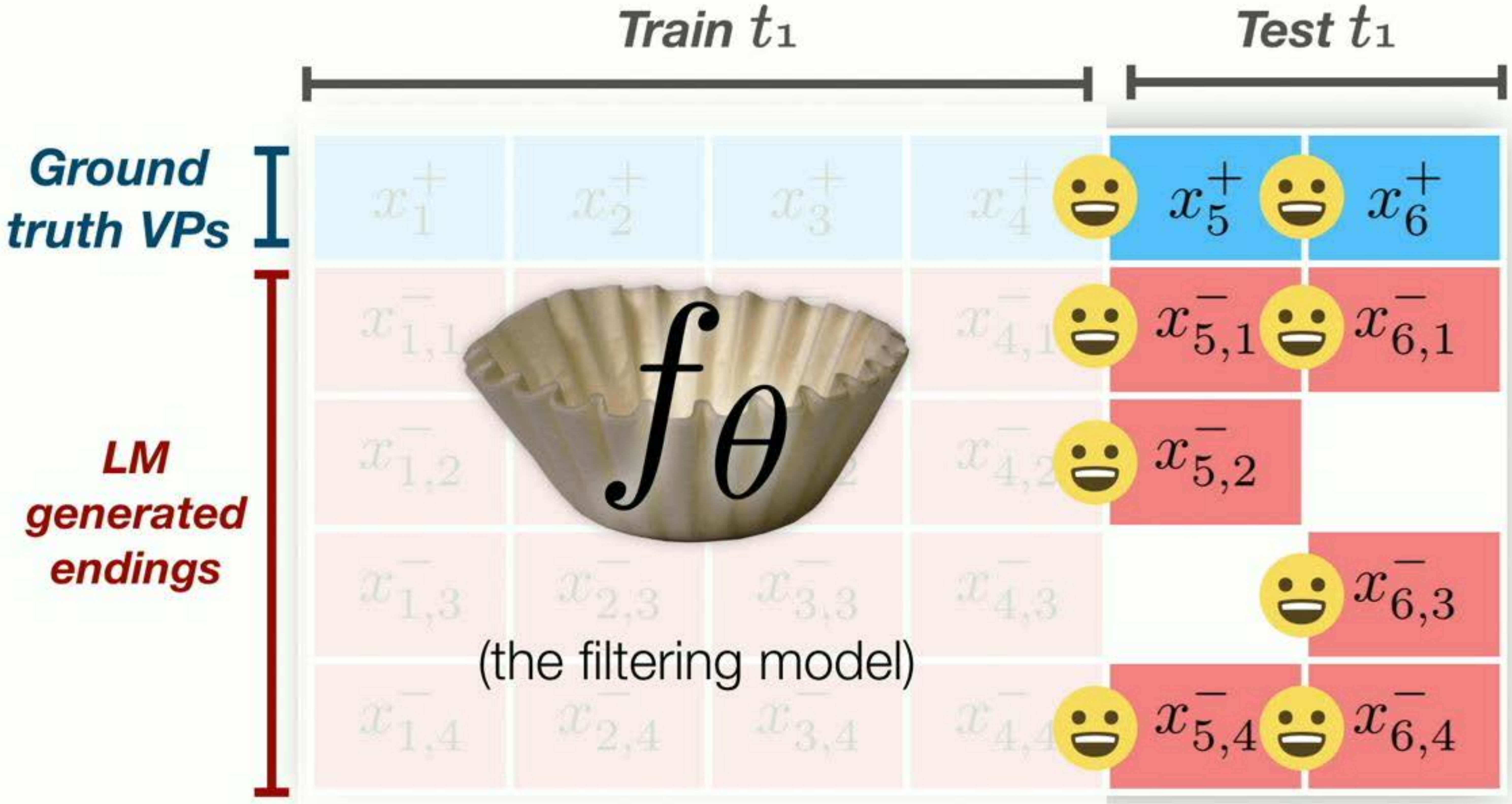

Train t_1
Test t_1
Ground truth VPs
LM generated endings
x_1^+ x_2^+ x_3^+ x_4^+ x_5^+ x_6^+
$x_{1,1}^-$ $x_{2,1}^-$ $x_{3,1}^-$ $x_{4,1}^-$ $x_{5,1}^-$ $x_{6,1}^-$
$x_{1,2}^-$ $x_{2,2}^-$ $x_{3,2}^-$ $x_{4,2}^-$ $x_{5,2}^-$ $x_{6,2}^-$
$x_{1,3}^-$ $x_{2,3}^-$ $x_{3,3}^-$ $x_{4,3}^-$ $x_{5,3}^-$ $x_{6,3}^-$
$x_{1,4}^-$ $x_{2,4}^-$ $x_{3,4}^-$ $x_{4,4}^-$ $x_{5,4}^-$ $x_{6,4}^-$



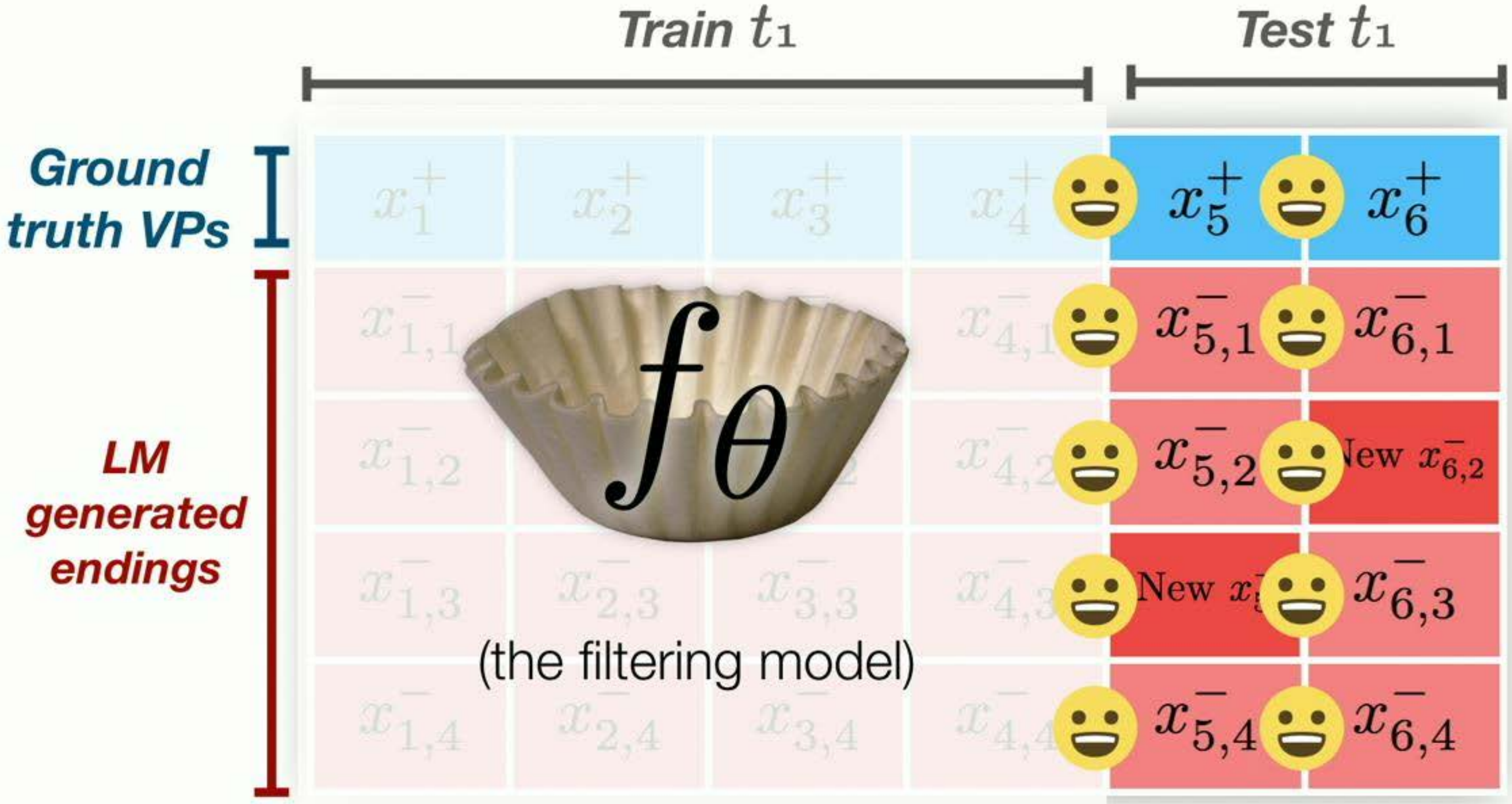
Train t_1
Test t_1
Ground truth VPs
LM generated endings
x_5^+ x_6^+
$x_{5,1}^-$ $x_{6,1}^-$
$x_{5,2}^-$ $x_{6,2}^-$
$x_{5,3}^-$ $x_{6,3}^-$
$x_{5,4}^-$ $x_{6,4}^-$
f_θ
(the filtering model)

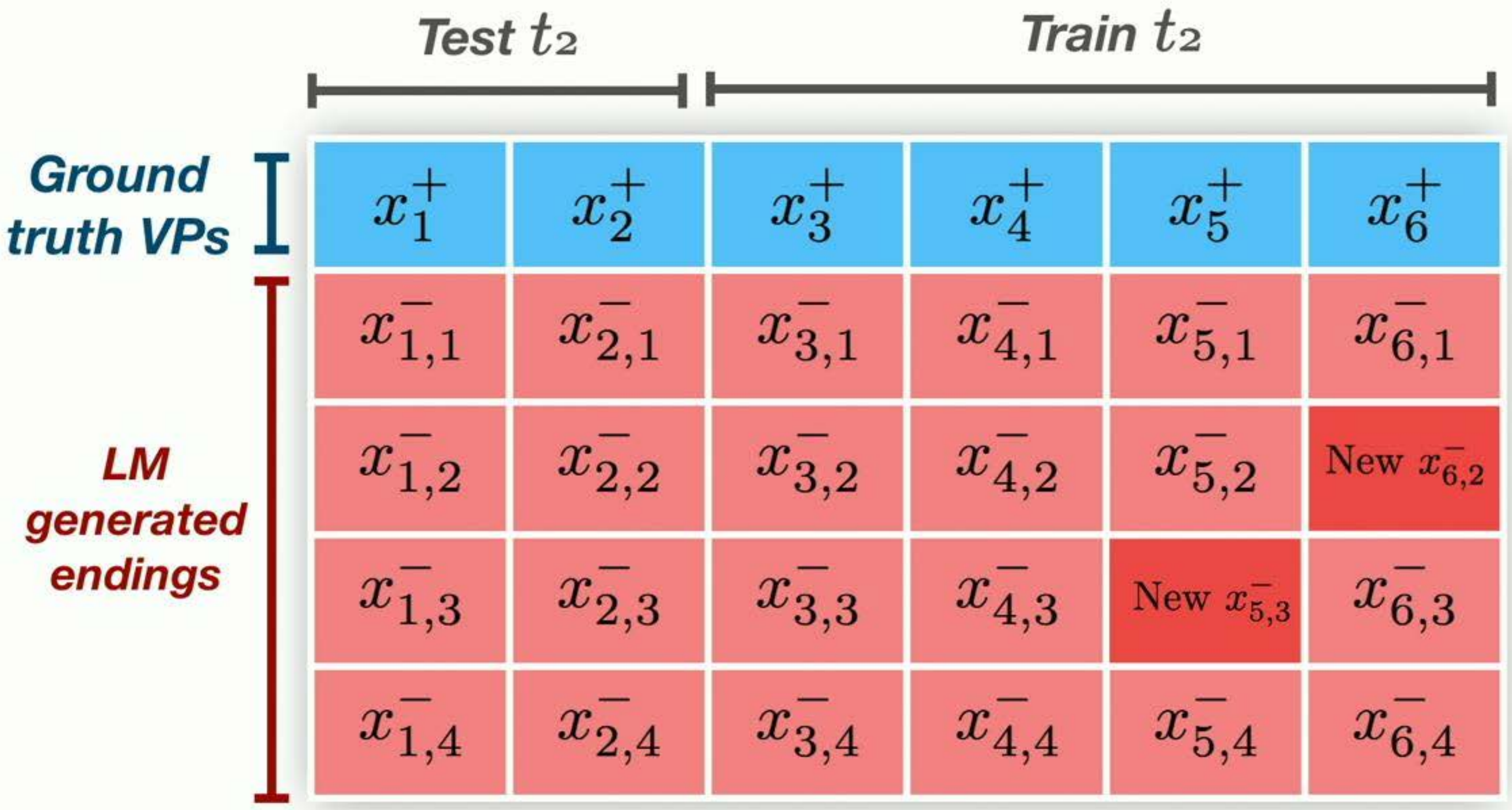


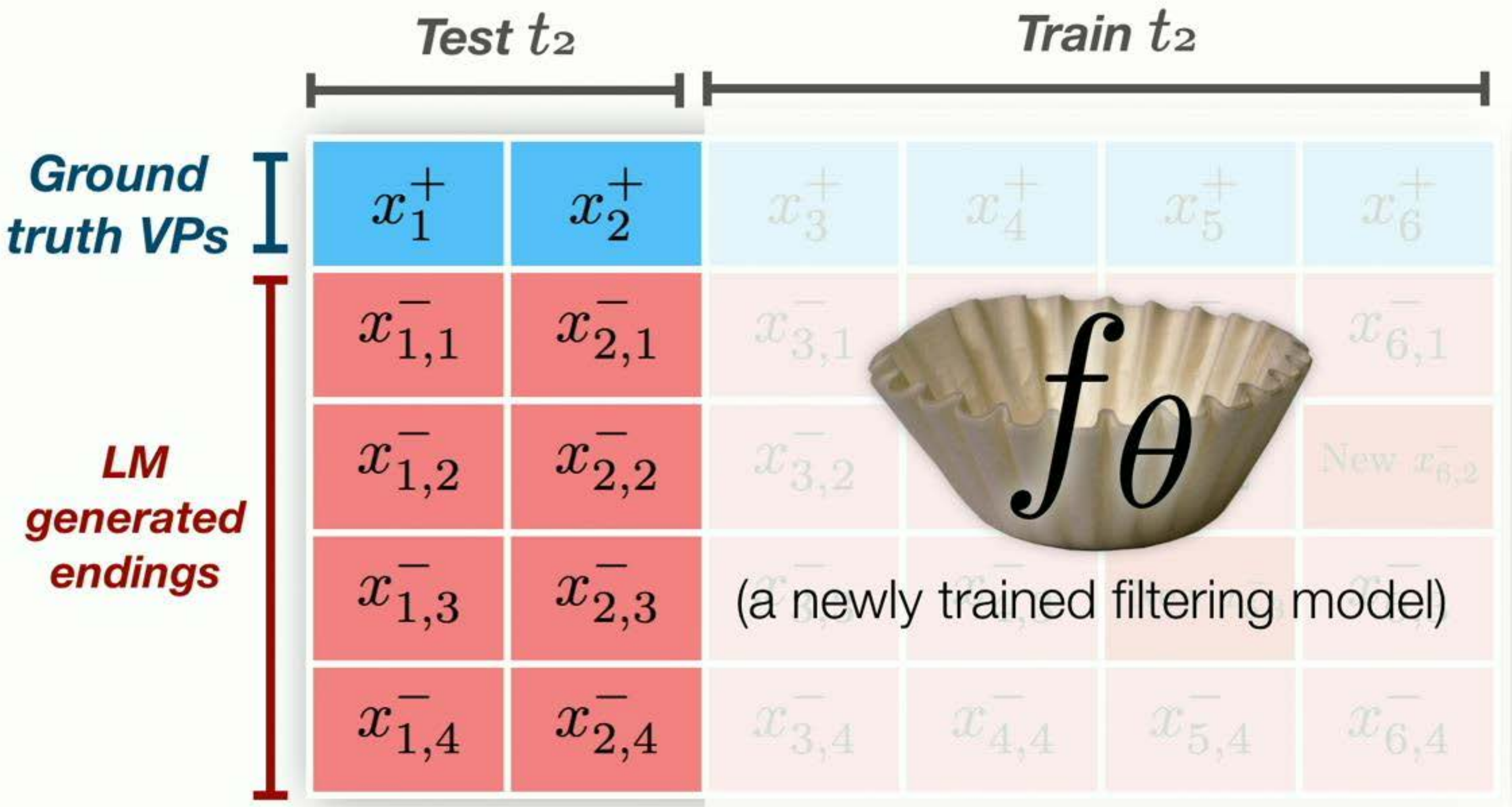
Train t_1
Test t_1
Ground truth VPs
LM generated endings
f_θ
(the filtering model)
x_1^+ x_2^+ x_3^+ x_4^+ x_5^+ x_6^+
$x_{1,1}^-$ $x_{4,1}^-$ $x_{5,1}^-$ $x_{6,1}^-$
$x_{1,2}^-$ $x_{4,2}^-$ $x_{5,2}^-$ $x_{6,2}^-$
$x_{1,3}^-$ $x_{2,3}^-$ $x_{3,3}^-$ $x_{4,3}^-$ $x_{5,3}^-$ $x_{6,3}^-$
$x_{1,4}^-$ $x_{2,4}^-$ $x_{3,4}^-$ $x_{4,4}^-$ $x_{5,4}^-$ $x_{6,4}^-$



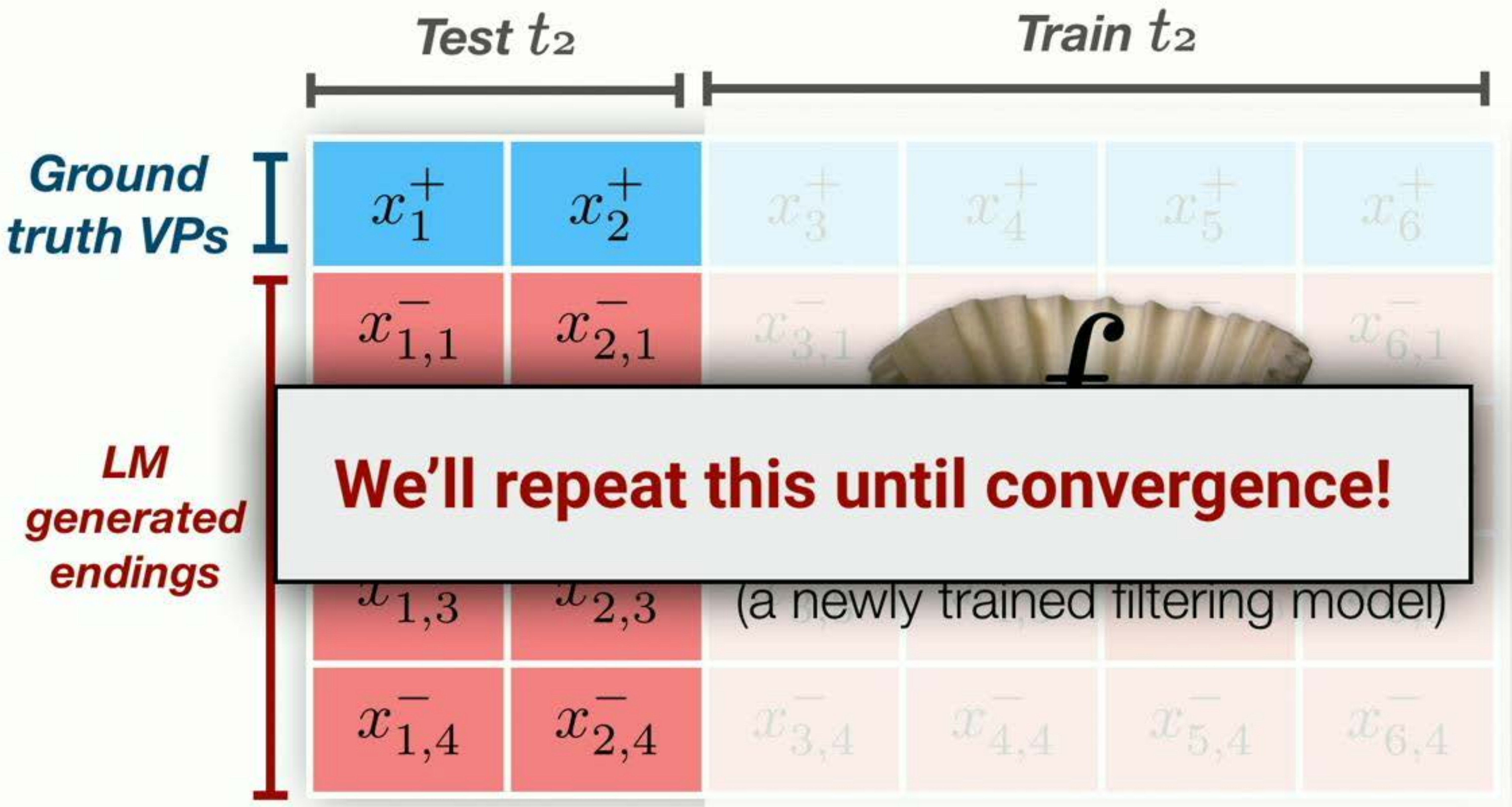
Train t_1
Test t_1
Ground truth VPs
LM generated endings
f_θ
(the filtering model)
x_1^+ x_2^+ x_3^+ x_4^+ x_5^+ x_6^+
$x_{1,1}^-$ $x_{4,1}^-$ $x_{5,1}^-$ $x_{6,1}^-$
$x_{1,2}^-$ $x_{4,2}^-$ $x_{5,2}^-$
$x_{1,3}^-$ $x_{2,3}^-$ $x_{3,3}^-$ $x_{4,3}^-$ $x_{6,3}^-$
$x_{1,4}^-$ $x_{2,4}^-$ $x_{3,4}^-$ $x_{4,4}^-$ $x_{5,4}^-$ $x_{6,4}^-$


Train t_1
Test t_1
Ground truth VPs
LM generated endings
x_1^+ x_2^+ x_3^+ x_4^+ x_5^+ x_6^+
$x_{1,1}^-$ $x_{4,1}^-$ $x_{5,1}^-$ $x_{6,1}^-$
$x_{1,2}^-$ $x_{4,2}^-$ $x_{5,2}^-$
New $x_{6,2}^-$
$x_{1,3}^-$ $x_{2,3}^-$ $x_{3,3}^-$ $x_{4,3}^-$ $x_{6,3}^-$
$x_{1,4}^-$ $x_{2,4}^-$ $x_{3,4}^-$ $x_{4,4}^-$ $x_{5,4}^-$ $x_{6,4}^-$
f_θ
(the filtering model)


Test t_2
Train t_2
Ground truth VPs
LM generated endings
x_1^+ x_2^+ x_3^+ x_4^+ x_5^+ x_6^+
$x_{1,1}^-$ $x_{2,1}^-$ $x_{3,1}^-$ $x_{4,1}^-$ $x_{5,1}^-$ $x_{6,1}^-$
$x_{1,2}^-$ $x_{2,2}^-$ $x_{3,2}^-$ $x_{4,2}^-$ $x_{5,2}^-$ New $x_{6,2}^-$
$x_{1,3}^-$ $x_{2,3}^-$ $x_{3,3}^-$ $x_{4,3}^-$ New $x_{5,3}^-$ $x_{6,3}^-$
$x_{1,4}^-$ $x_{2,4}^-$ $x_{3,4}^-$ $x_{4,4}^-$ $x_{5,4}^-$ $x_{6,4}^-$



Test t_2
Train t_2
Ground truth VPs
LM generated endings
x_1^+ x_2^+ x_3^+ x_4^+ x_5^+ x_6^+
$x_{1,1}^-$ $x_{2,1}^-$ $x_{3,1}^-$ $x_{6,1}^-$
$x_{1,2}^-$ $x_{2,2}^-$ $x_{3,2}^-$ New $x_{6,2}^-$
$x_{1,3}^-$ $x_{2,3}^-$
$x_{1,4}^-$ $x_{2,4}^-$ $x_{3,4}^-$ $x_{4,4}^-$ $x_{5,4}^-$ $x_{6,4}^-$
f_θ
(a newly trained filtering model)


Test t_2
Train t_2
Ground truth VPs
LM generated endings
x_1^+
x_2^+
$x_{1,1}^-$
$x_{2,1}^-$
$x_{1,3}^-$
$x_{2,3}^-$
$x_{1,4}^-$
$x_{2,4}^-$
f
(a newly trained filtering model)
We'll repeat this until convergence!

What AF does and doesn't do

Somewhat stolen from Percy Liang

What AF does and doesn’t do

Train $\sim$ Test ?

Somewhat stolen from Percy Liang

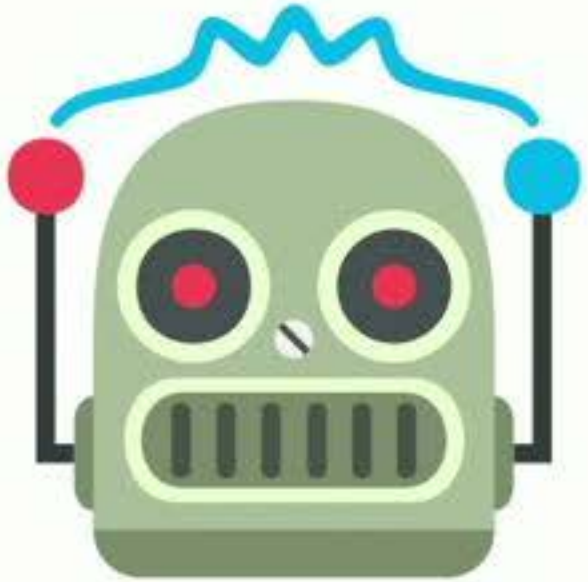
What AF does and doesn't do

Train ~ Test ?

Somewhat stolen from Percy Liang

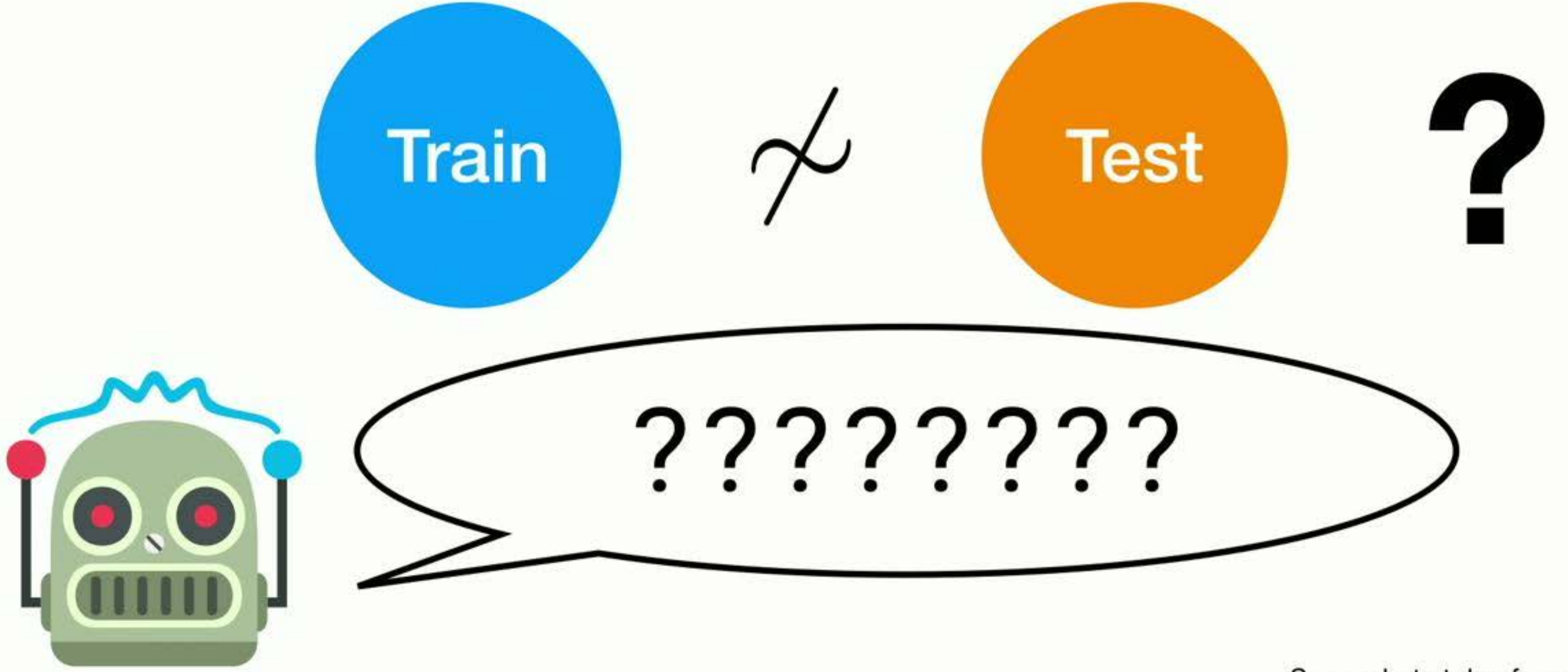
What AF does and doesn't do

Train $\nsim$ Test ?



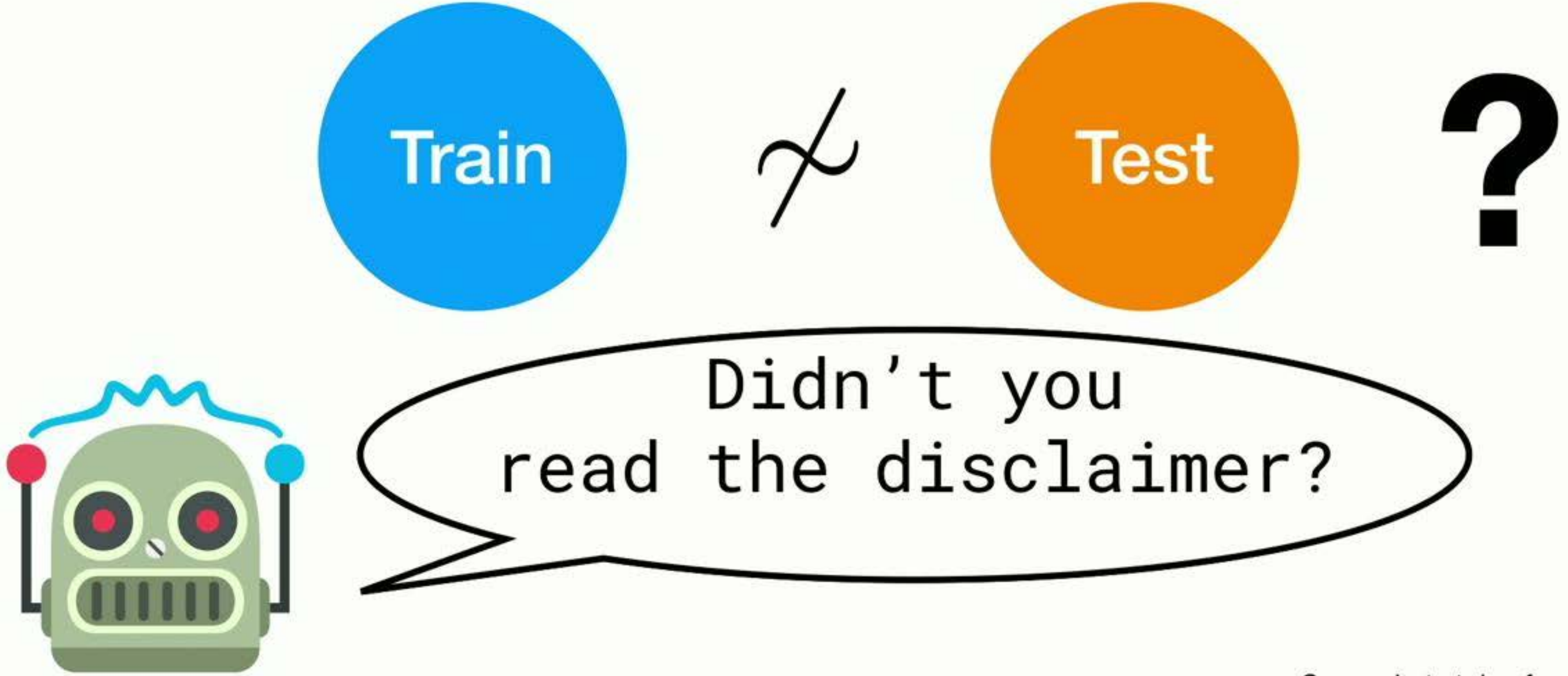
Somewhat stolen from Percy Liang

What AF does and doesn’t do



Somewhat stolen from Percy Liang

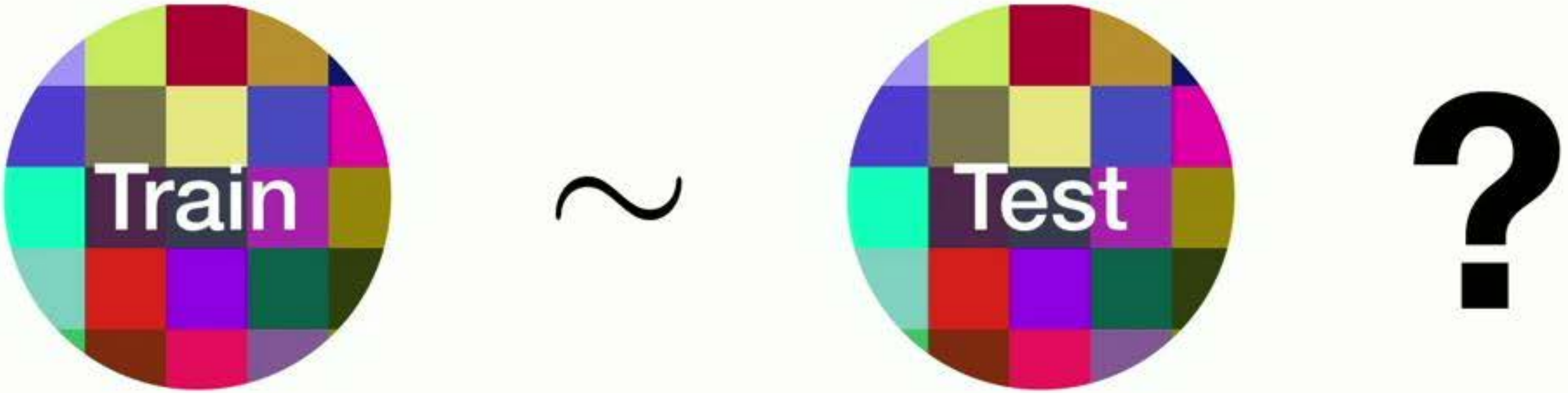
What AF does and doesn’t do



Somewhat stolen from Percy Liang

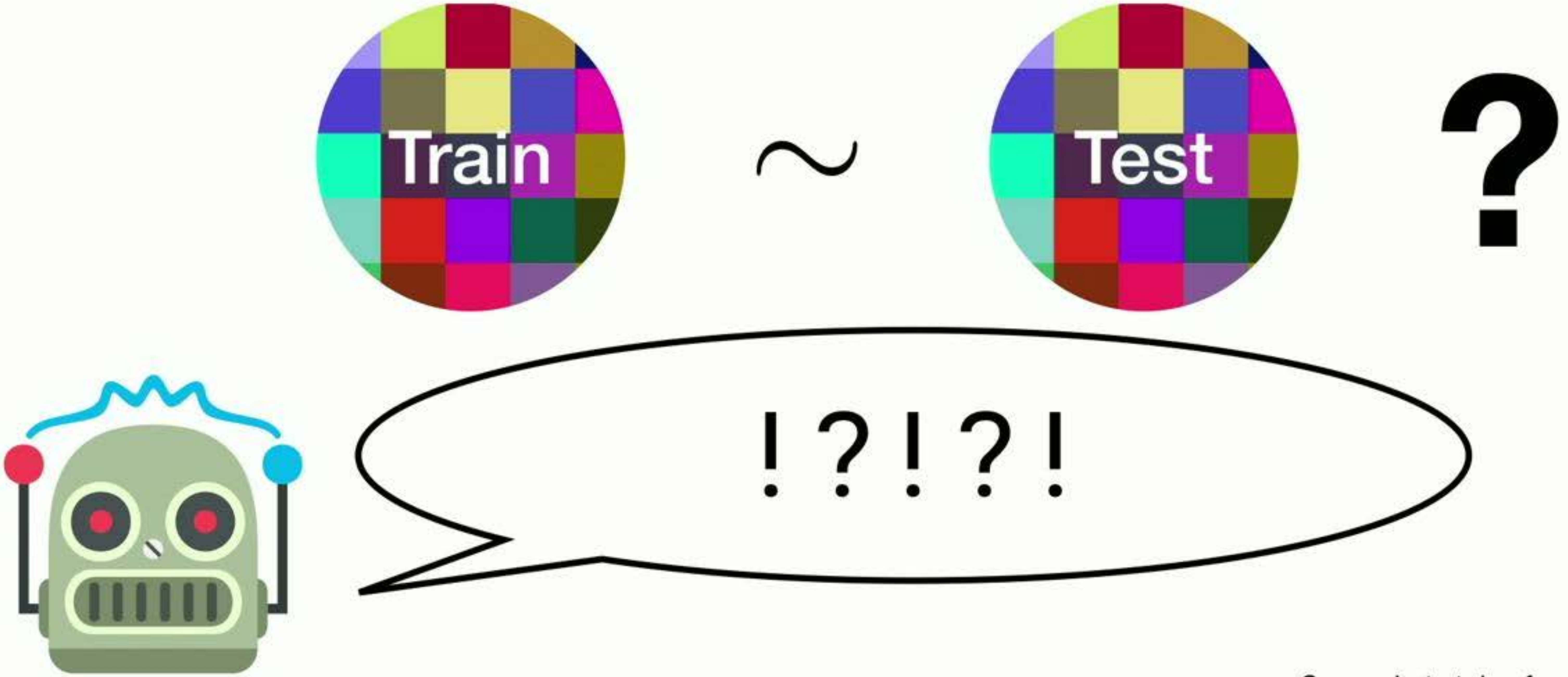
THIS MODEL SHOULD ONLY BE EVALUATED ON DATA THAT'S SIMILAR TO THE TRAINING SET!!!!

What AF does and doesn't do



Somewhat stolen from Percy Liang

What AF does and doesn’t do



Somewhat stolen from Percy Liang

What AF does and doesn’t do

Somewhat stolen from Percy Liang

Math behind AF

Dataset $\mathcal{D} = \{(x_i, y_i)\}_{1 \leq i \leq N}$ Model $f_\theta : \mathcal{X} \to \mathbb{R}^{|\mathcal{Y}|}$

Loss function $L(f_\theta, \mathcal{D})$

Math behind AF

Dataset $\mathcal{D} = \{(x_i, y_i)\}_{1 \leq i \leq N}$ Model $f_\theta : \mathcal{X} \to \mathbb{R}^{|\mathcal{Y}|}$

Loss function $L(f_\theta, \mathcal{D})$

Empirical error

$$\boxed{\begin{gathered} I(\mathcal{D}, f) = \frac{1}{N} \sum_{i=1}^{N} L(f_{\theta_i^\star}, \{(x_i, y_i)\}), \\ \text{where } \theta_i^\star = \operatorname*{argmin}_\theta L(f_\theta, \mathcal{D} \setminus \{(x_i, y_i)\}), \end{gathered}}$$

Math behind AF

Dataset $\mathcal{D} = \{(x_i, y_i)\}_{1 \le i \le N}$ Model $f_\theta : \mathcal{X} \to \mathbb{R}^{|\mathcal{Y}|}$

Loss function $L(f_\theta, \mathcal{D})$

Empirical error

$$I(\mathcal{D}, f) = \frac{1}{N} \sum_{i=1}^{N} L(f_{\theta_i^\star}, \{(x_i, y_i)\}),$$

$$\text{where } \theta_i^\star = \underset{\theta}{\operatorname{argmin}}\, L(f_\theta, \mathcal{D} \setminus \{(x_i, y_i)\}),$$

The best parameters as determined by other datapoints

Math behind AF

Dataset $\mathcal{D} = \{(x_i, y_i)\}_{1 \leq i \leq N}$ Model $f_\theta : \mathcal{X} \to \mathbb{R}^{|\mathcal{Y}|}$

How much does knowing N-1 datapoints help?

Empirical error

$$I(\mathcal{D}, f) = \frac{1}{N} \sum_{i=1}^{N} L(f_{\theta_i^\star}, \{(x_i, y_i)\}),$$

$$\text{where } \theta_i^\star = \operatorname*{argmin}_\theta L(f_\theta, \mathcal{D} \setminus \{(x_i, y_i)\}),$$

The best parameters as determined by other datapoints

Math behind AF

D … $|\mathcal{Y}|$

Let's MAXIMIZE dataset-level empirical error!

How much does knowing N-1 datapoints help?

Empirical error

$$I(\mathcal{D}, f) = \frac{1}{N} \sum_{i=1}^{N} L(f_{\theta_i^\star}, \{(x_i, y_i)\}),$$

$$\text{where } \theta_i^\star = \operatorname*{argmin}_{\theta} L(f_\theta, \mathcal{D} \setminus \{(x_i, y_i)\}),$$

The best parameters as determined by other datapoints

Adversarial Filters for SWAG

We used an ensemble of stylistic models centered around the ending sentence. These are powerful adversaries!
(Schwartz et al., 2017, Gururangan et al., 2018)

Adversarial Filters for SWAG

We used an ensemble of stylistic models centered around the ending sentence. These are powerful adversaries!
(Schwartz et al., 2017, Gururangan et al., 2018)

A multilayer perceptron, given LM perplexities as features

Bag-of-words

CNN

BiLSTM, with uncommon words replaced by POS tags

Adversarial Filters for SWAG

We used an ensemble of stylistic models centered around the ending sentence. These are powerful adversaries!
(Schwartz et al., 2017, Gururangan et al., 2018)

After 100 iterations of adversarial filtering, the accuracy of these stylistic models drops to chance.

Adversarial Filters for SWAG

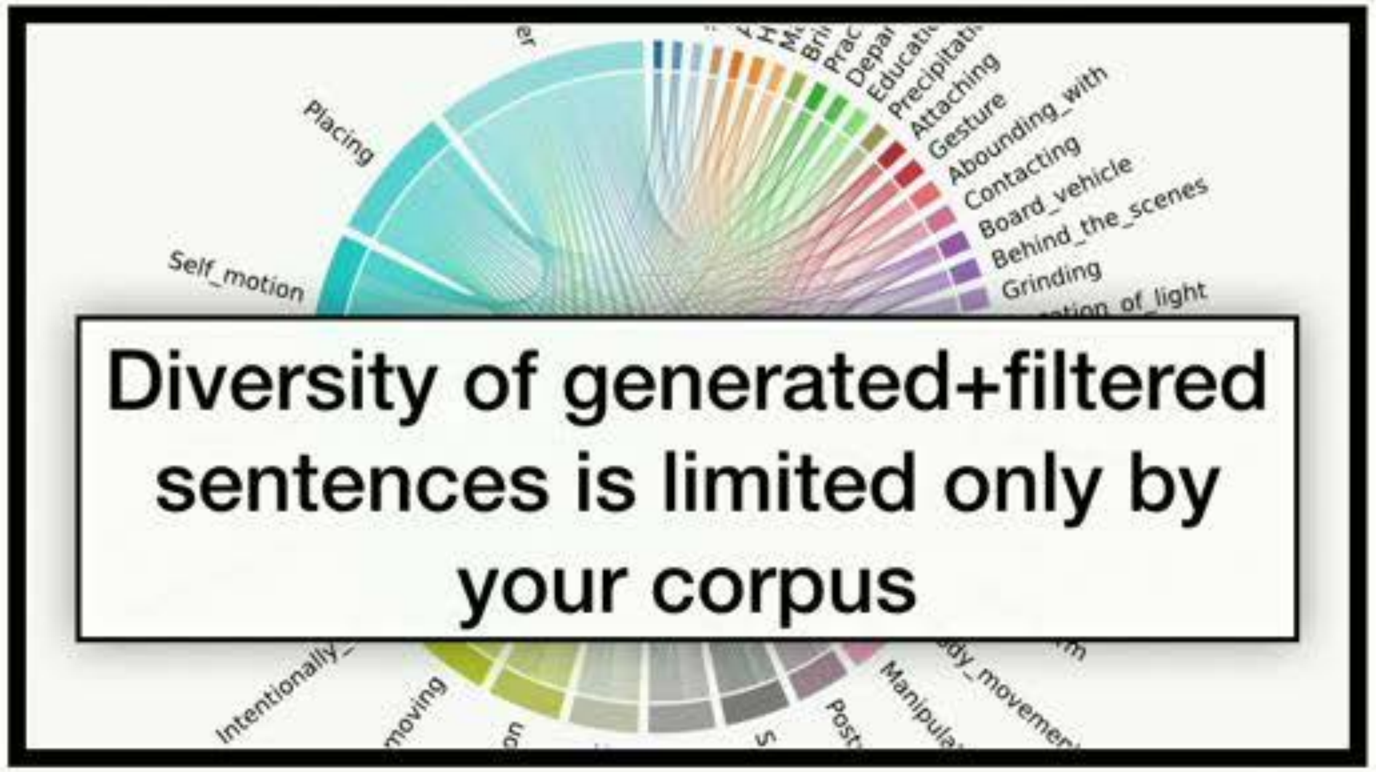
We used an ensemble of stylistic models centered around the ending sentence. These are powerful adversaries!
(Schwartz et al., 2017, Gururangan et al., 2018)

After 100 iterations of adversarial filtering, the accuracy of these stylistic models drops to chance.

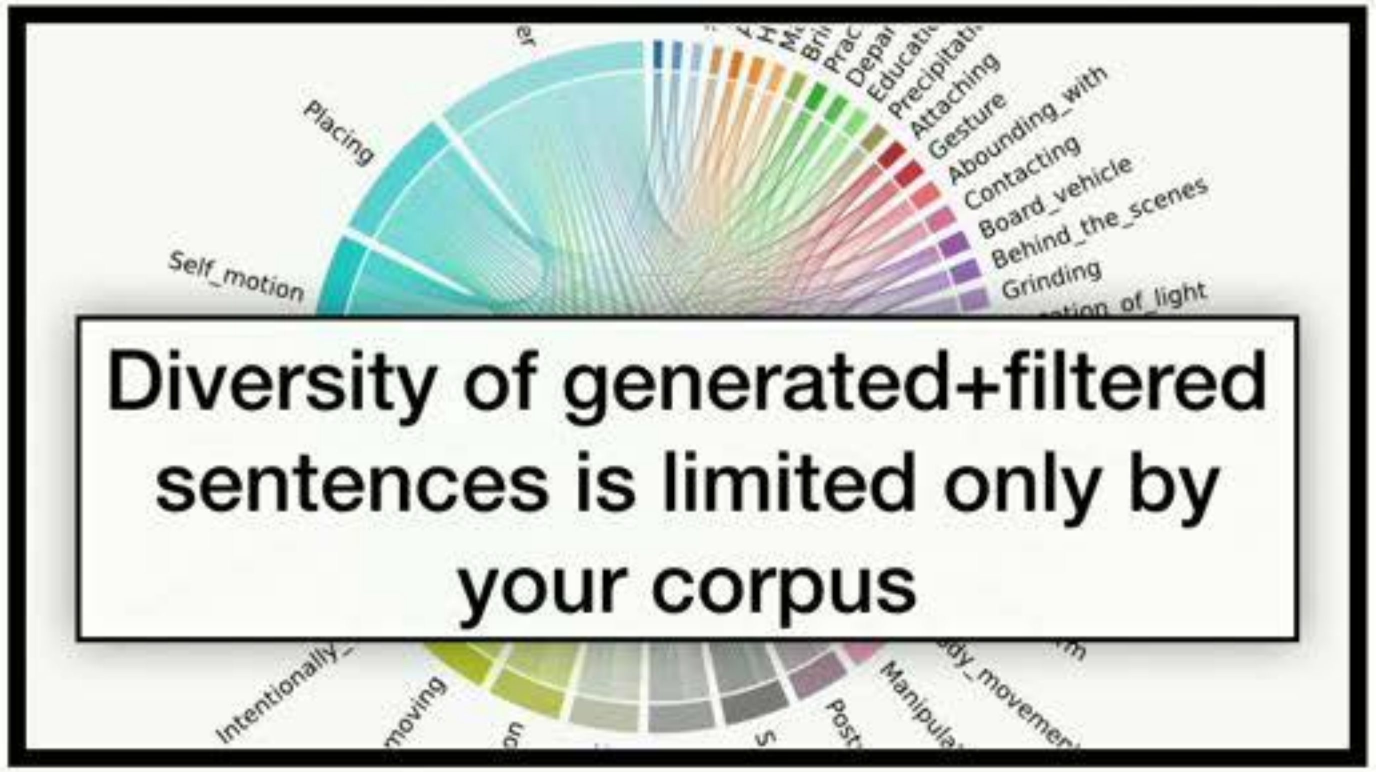
Last, crowd workers validate the entire dataset to remove false negatives, leaving us with 110k multiple-choice questions.

Unique Contributions of Adversarial Filtering

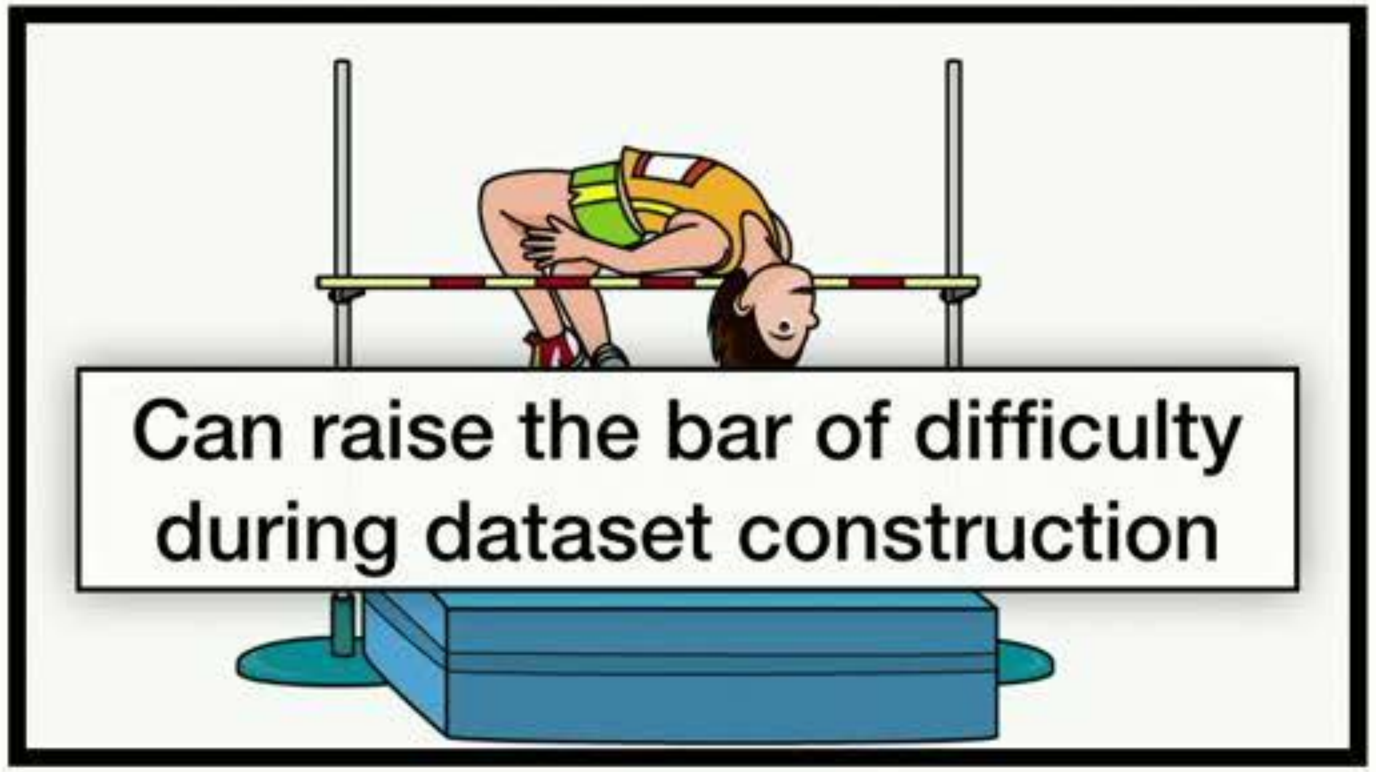
Unique Contributions of Adversarial Filtering



Unique Contributions of Adversarial Filtering

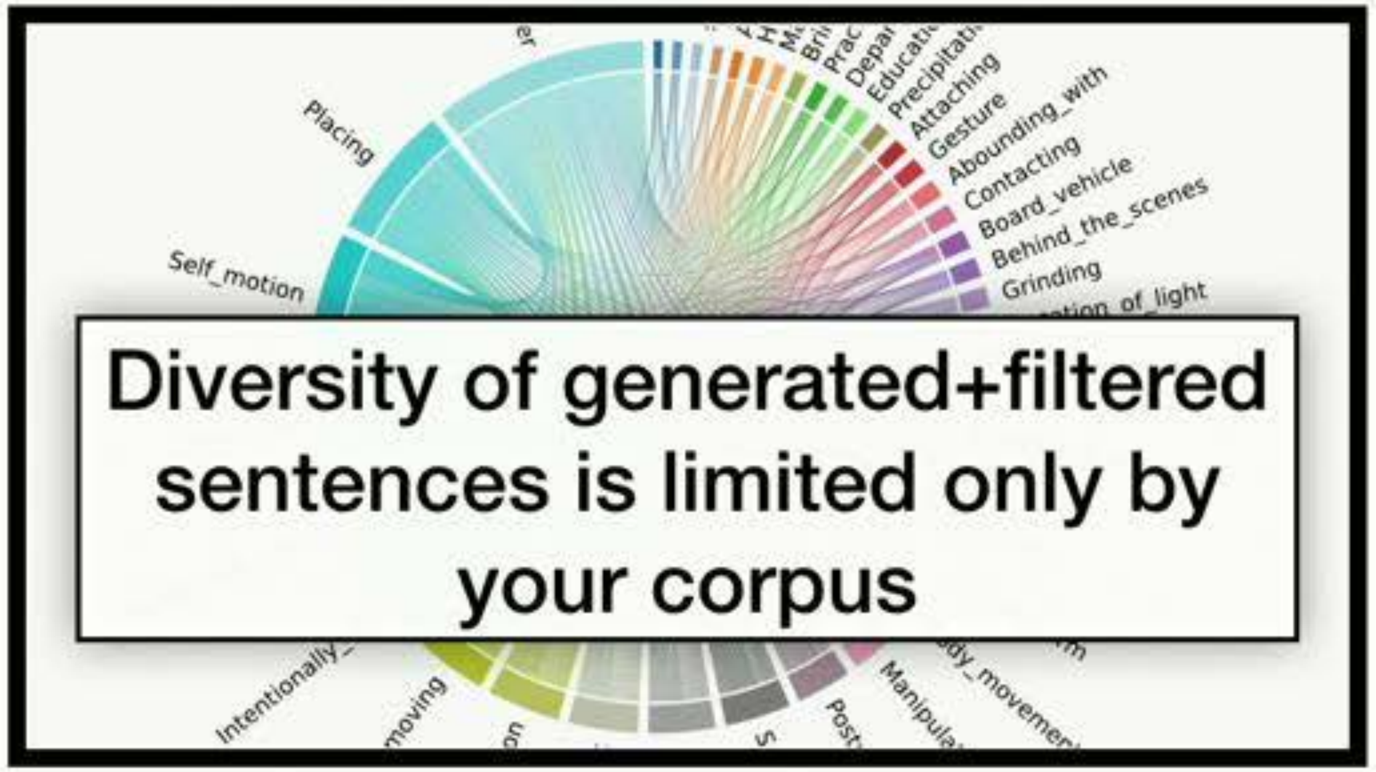


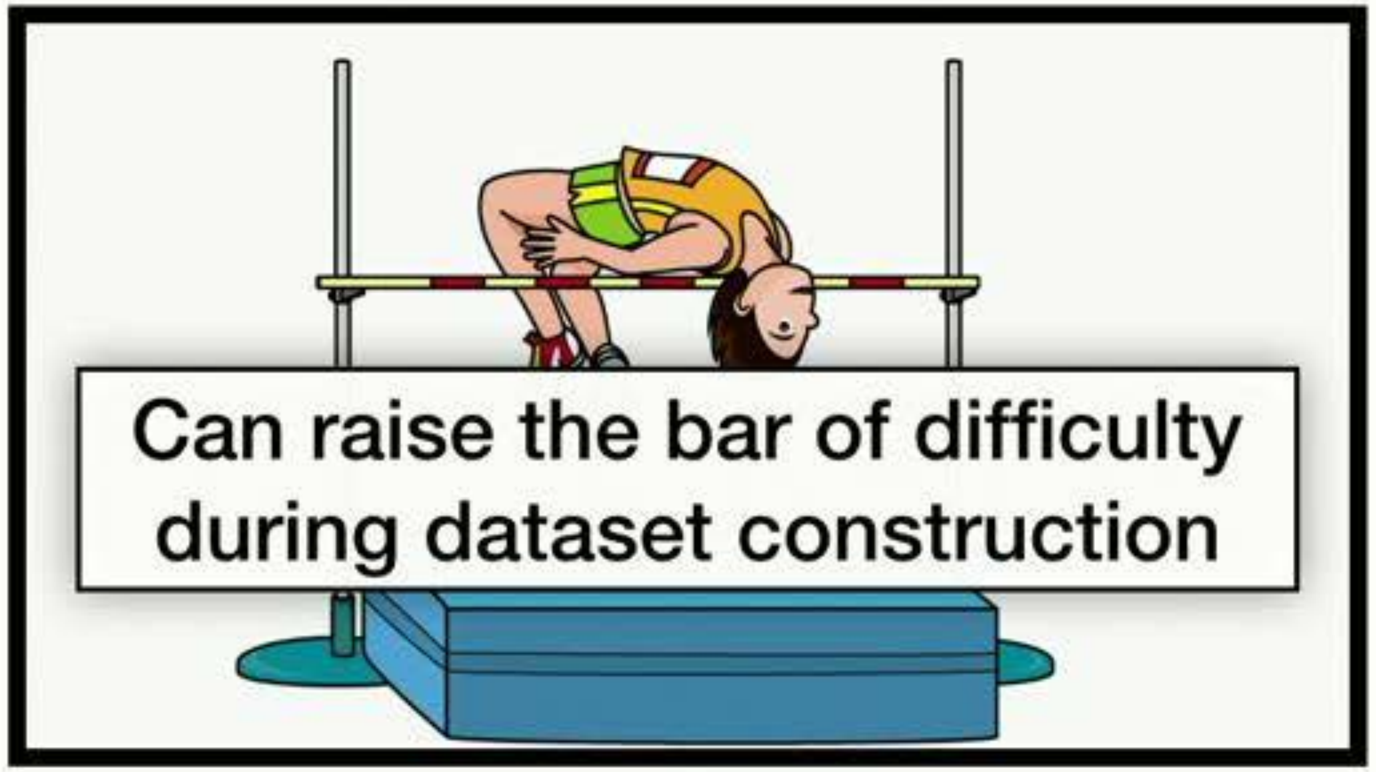
Diversity of generated+filtered sentences is limited only by your corpus

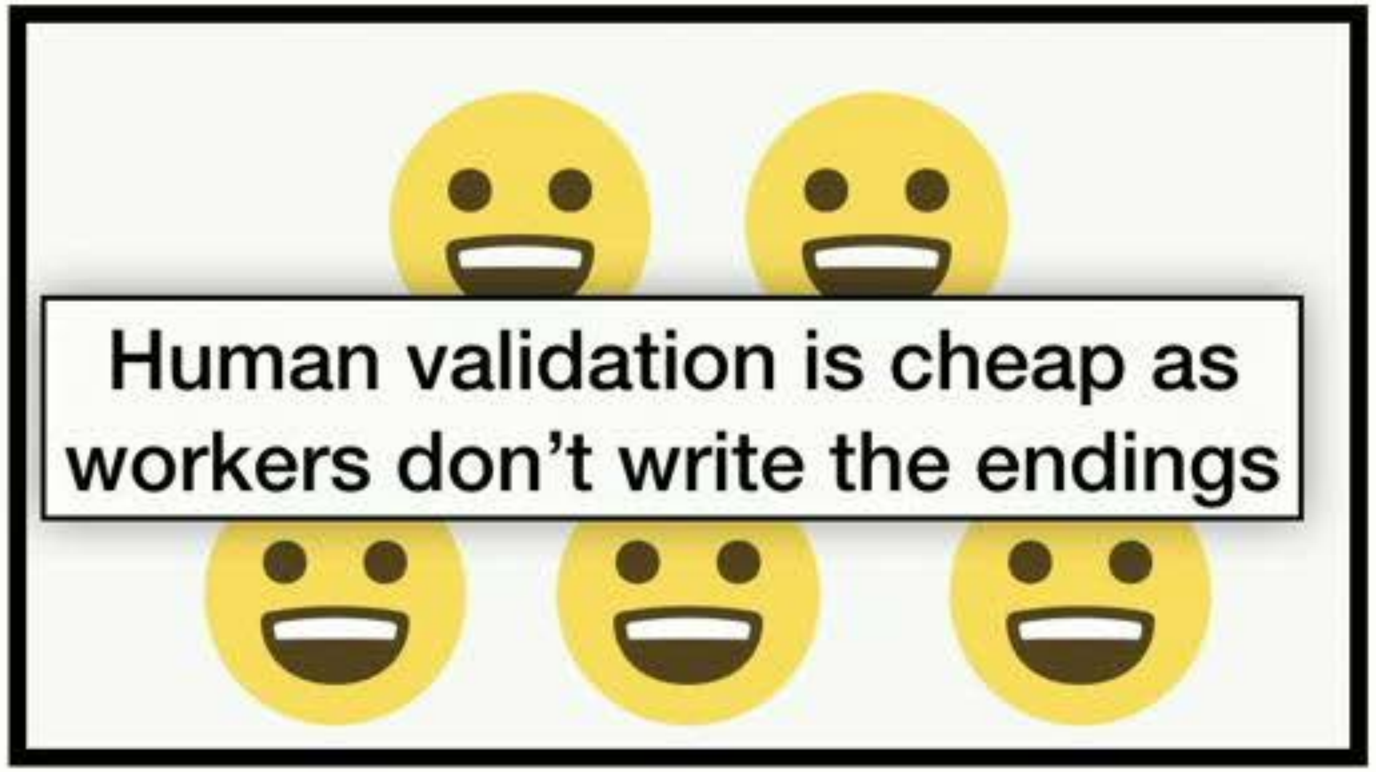


Can raise the bar of difficulty during dataset construction

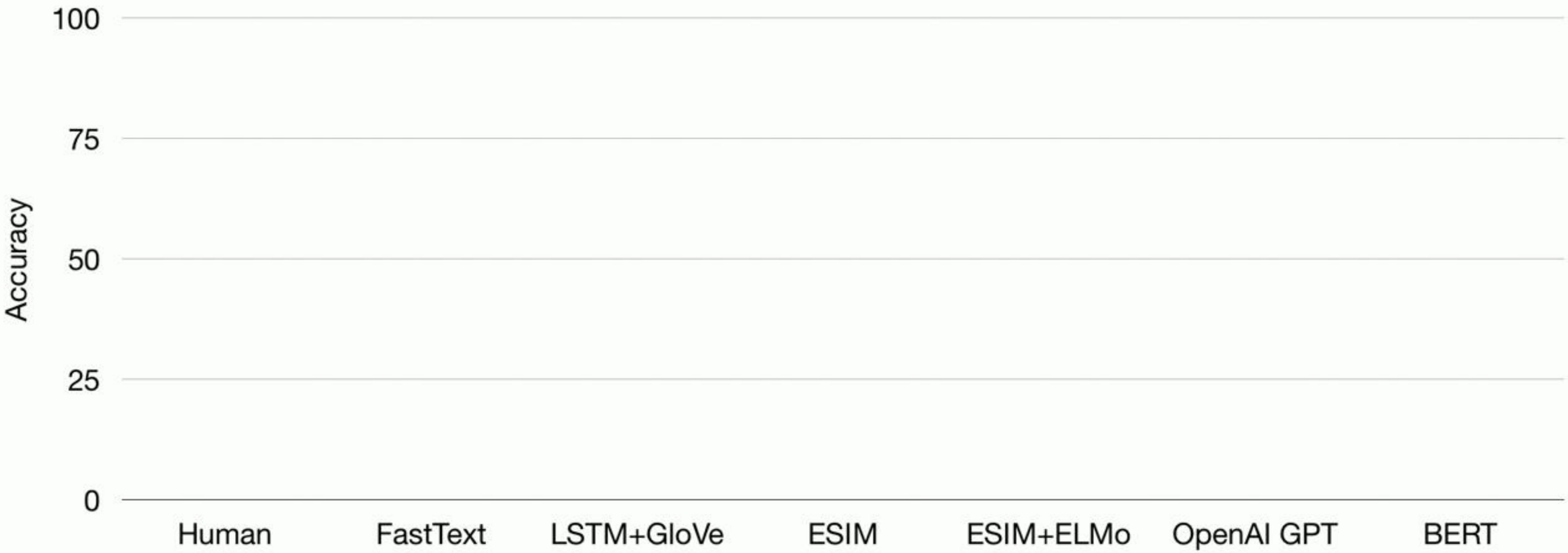
Unique Contributions of Adversarial Filtering



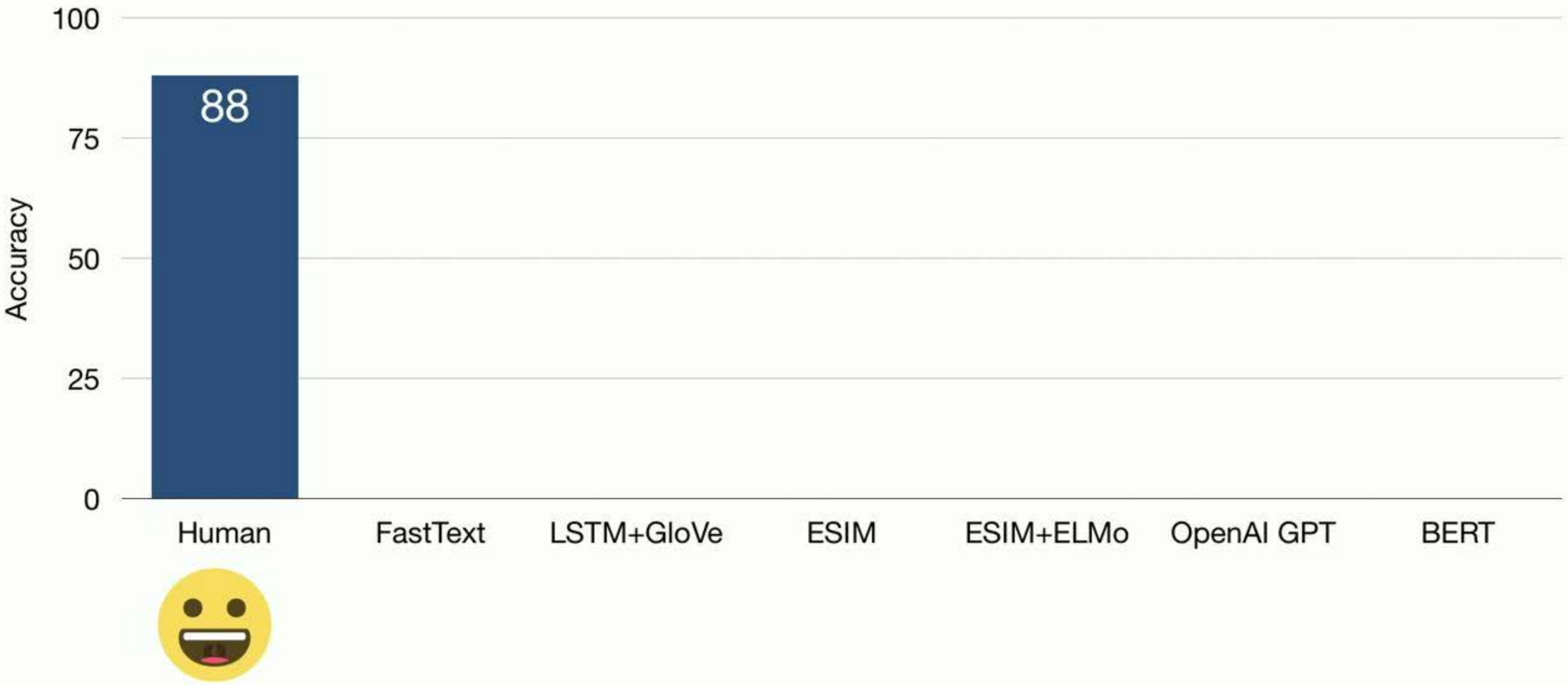




SWAG Results

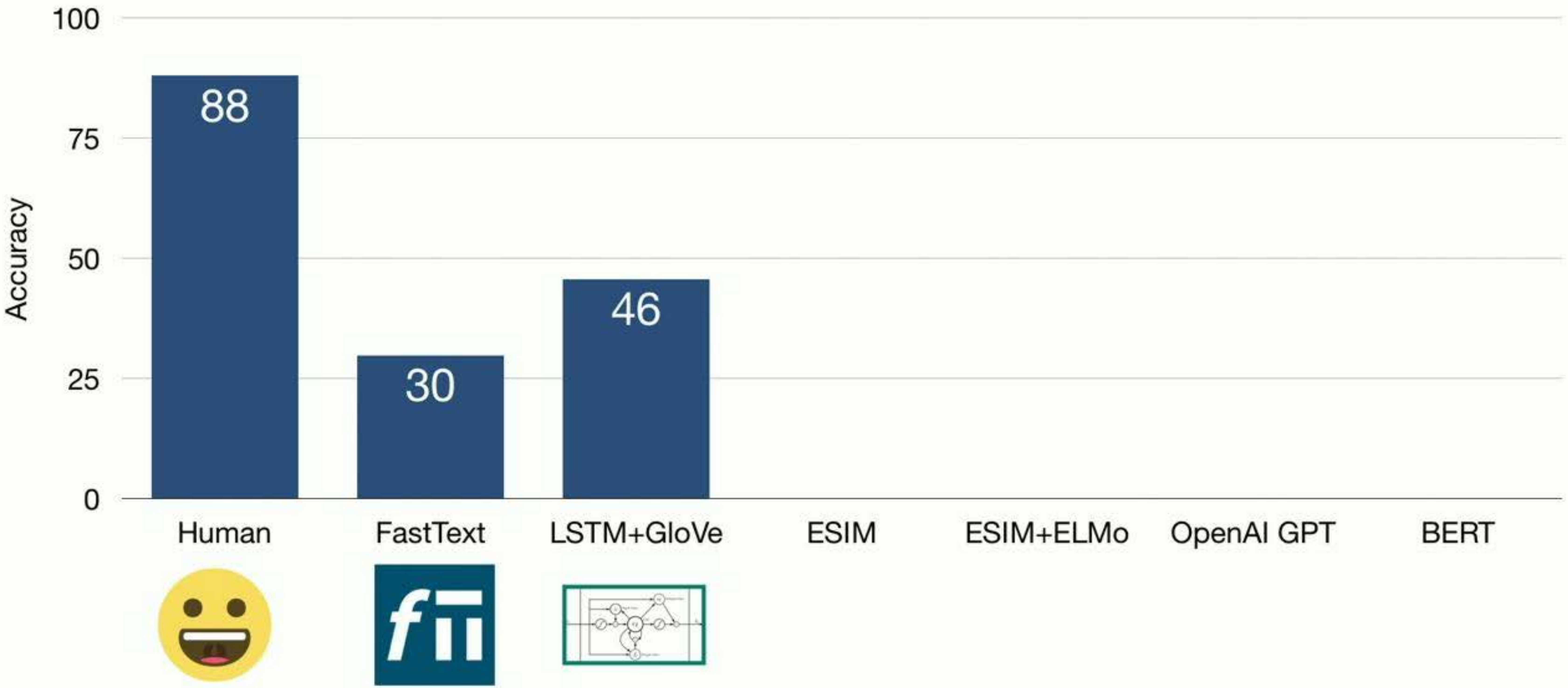


SWAG Results

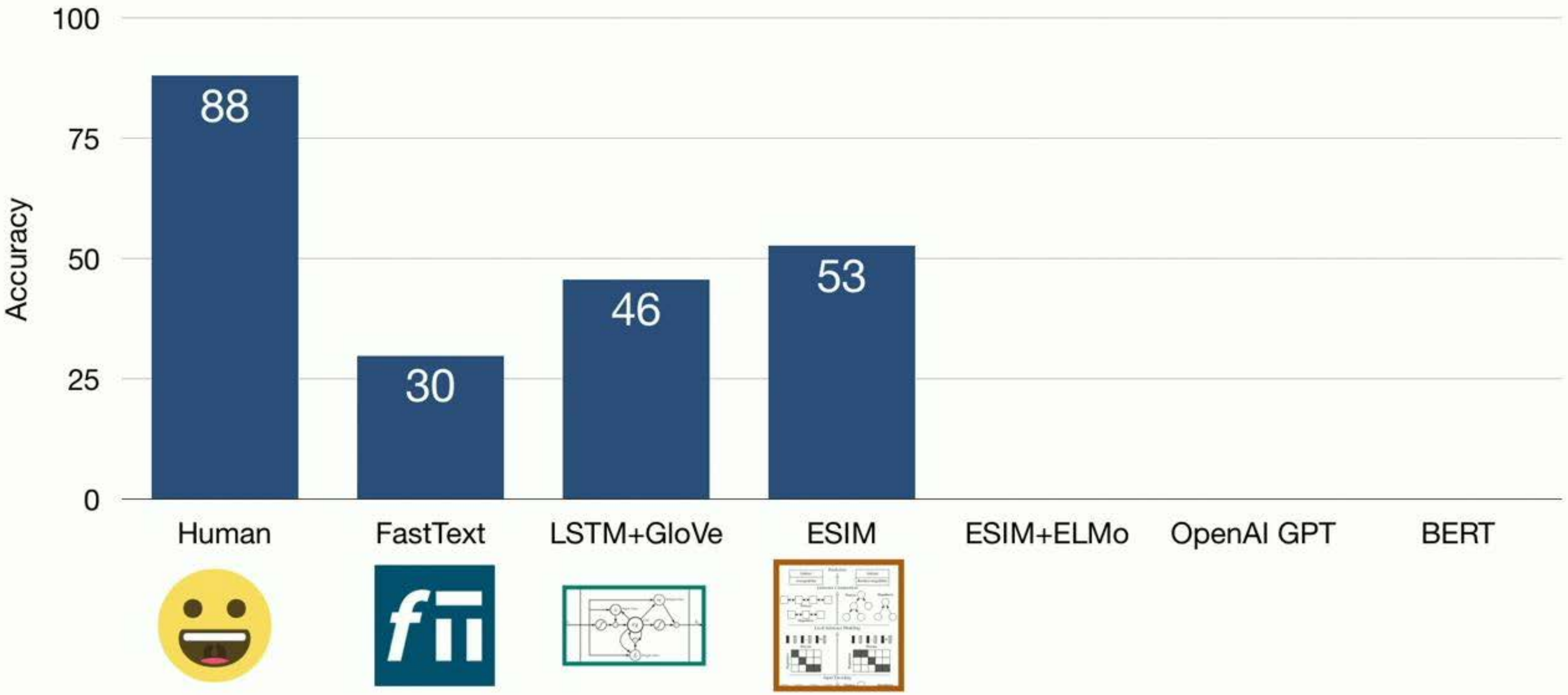


SWAG Results

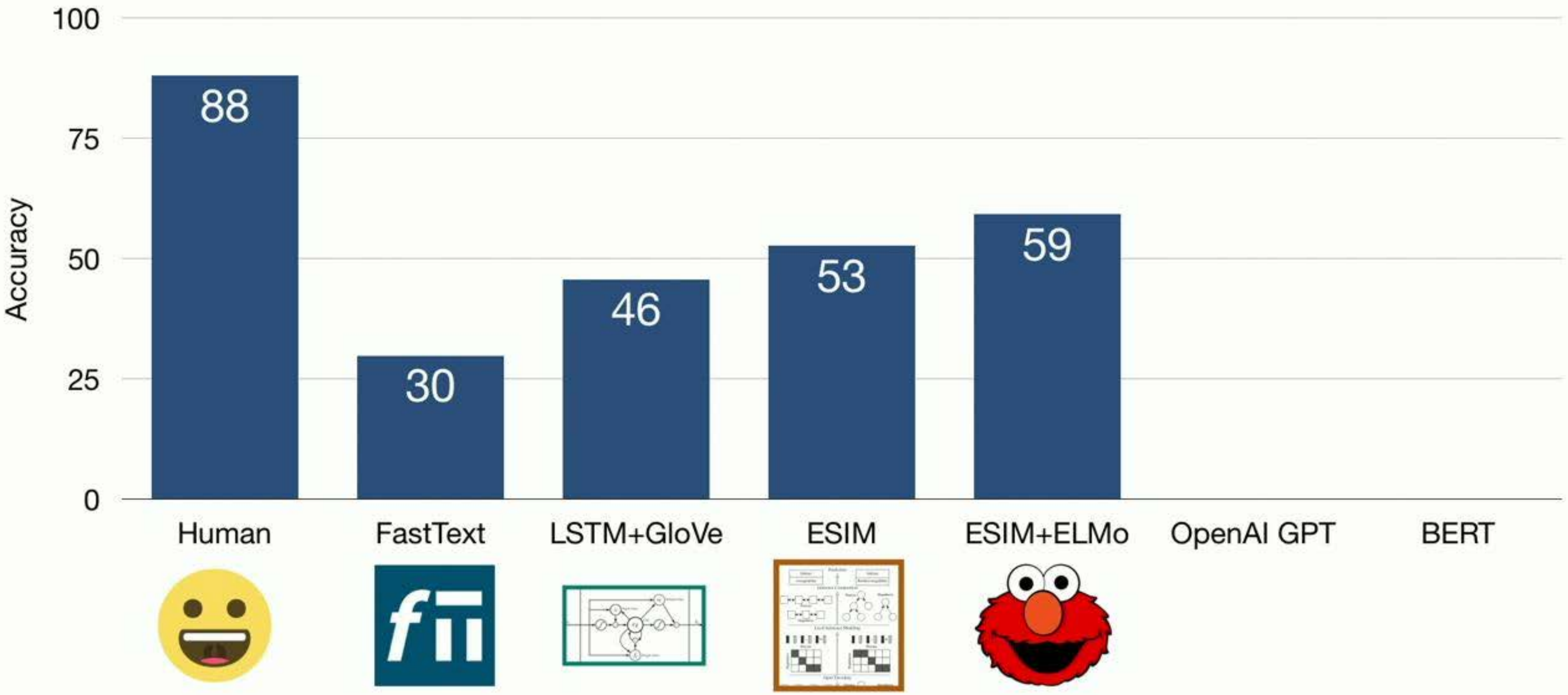
SWAG Results



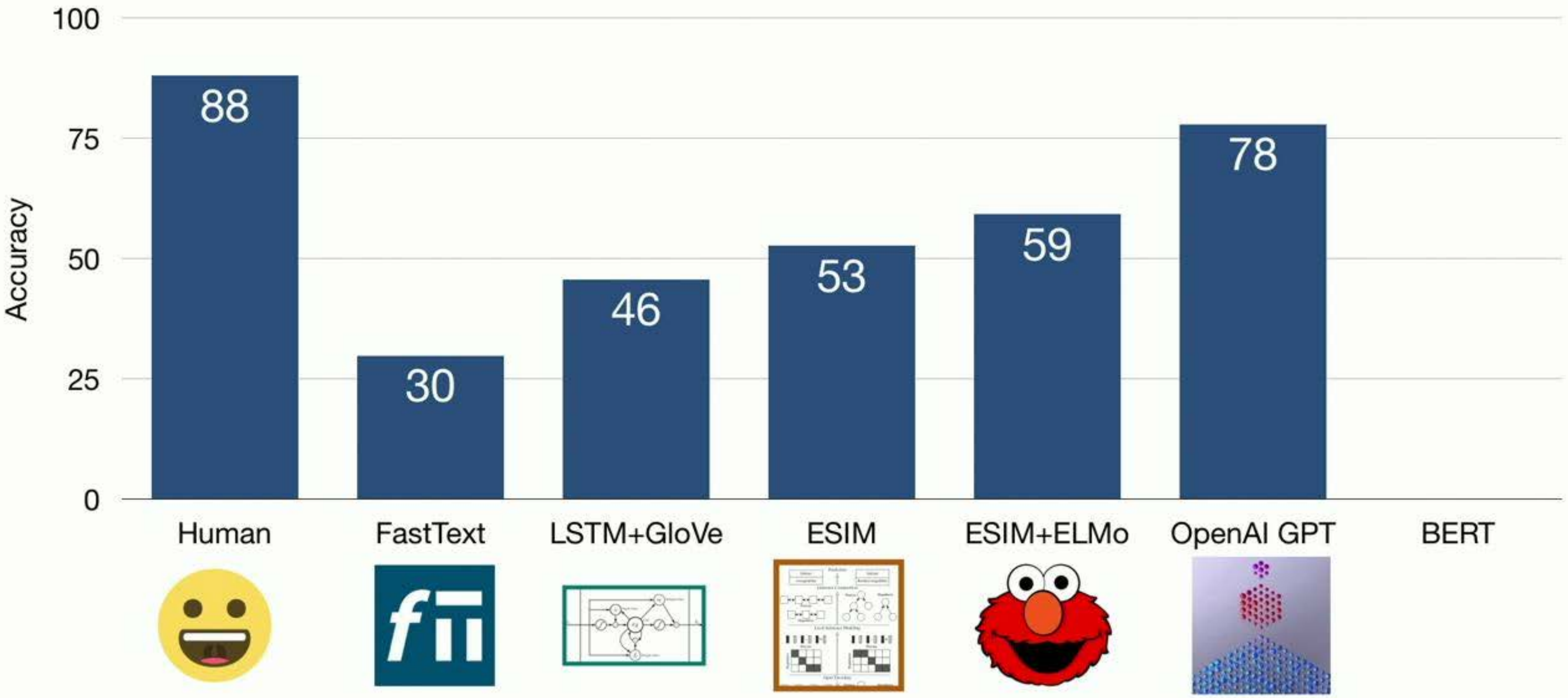
SWAG Results



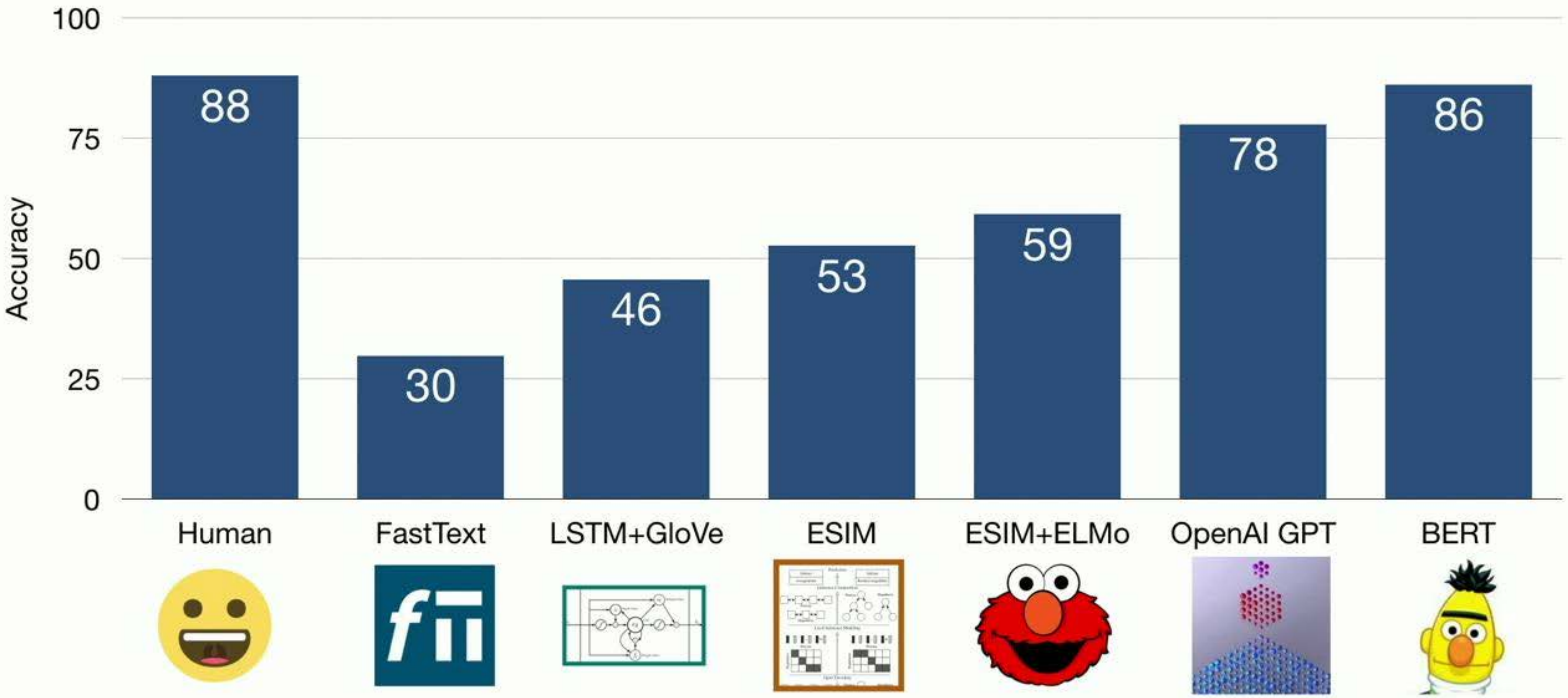
SWAG Results



SWAG Results

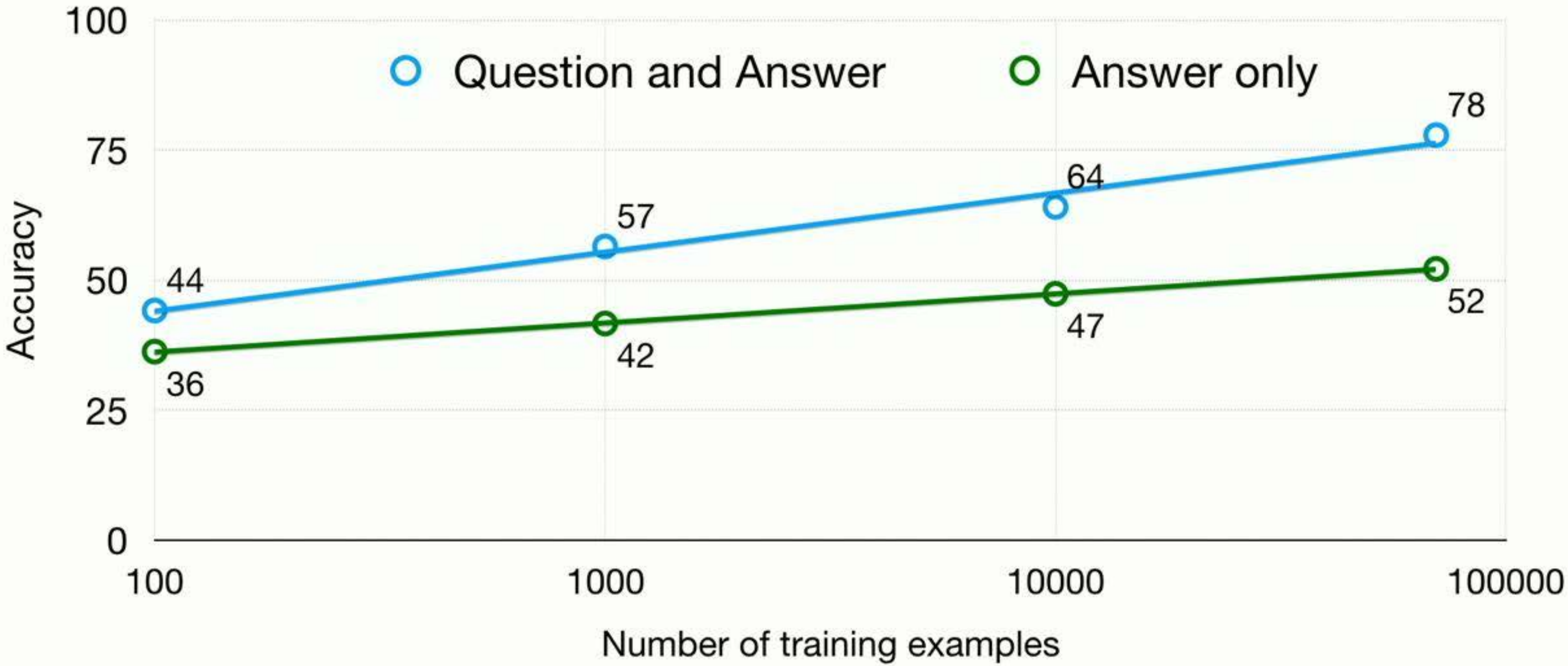


SWAG Results



What’s next?

OpenAI GPT performance vs. training set size



What do models learn by fine-tuning on SWAG?

A girl is going across the monkey bars.

She gets to the other side and stands on a wooden plank.

What do models learn by fine-tuning on SWAG?

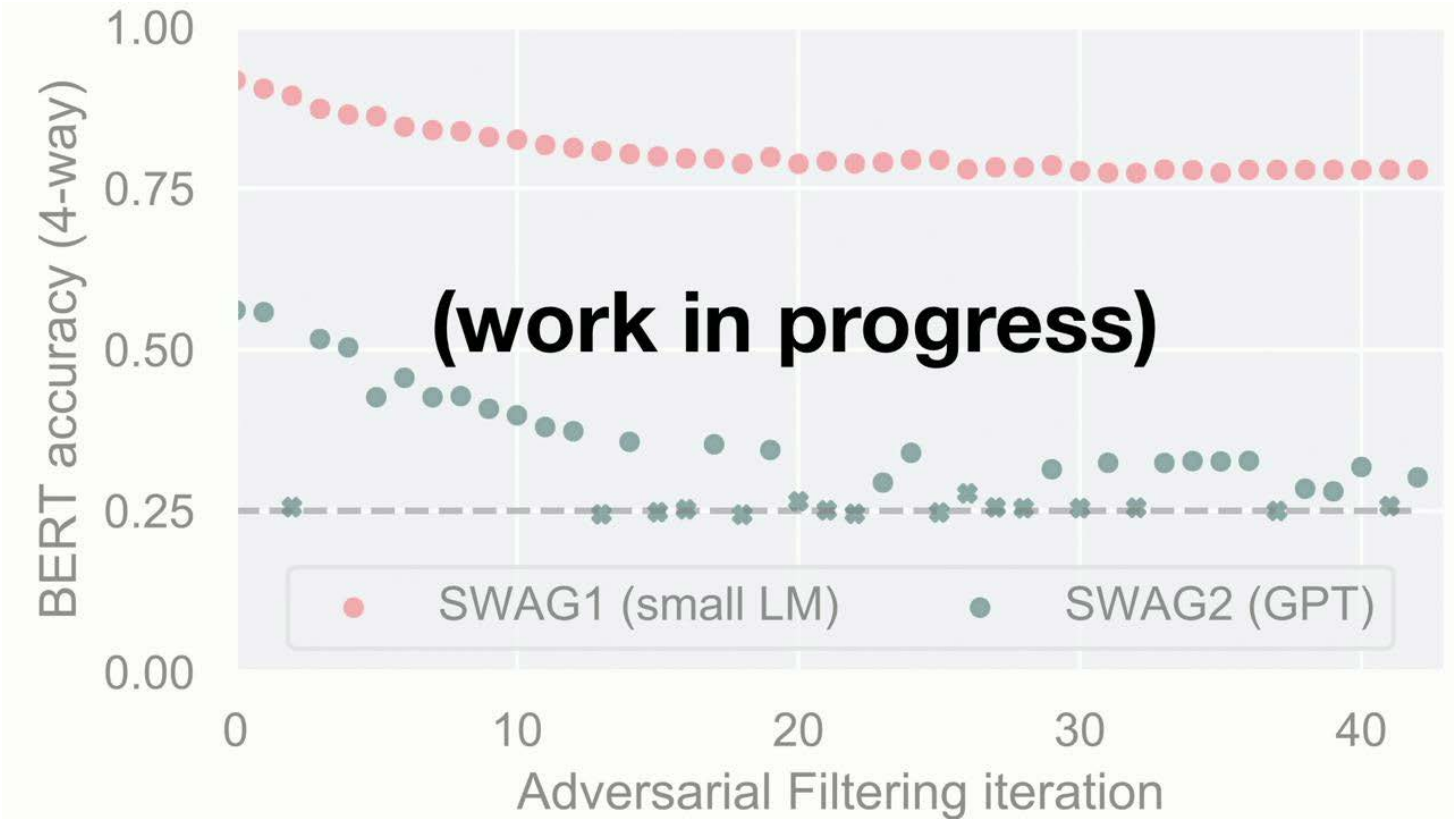
A girl is going across the monkey bars.

She gets to the other side and stands on a wooden plank.

Can a 2 sentence NLI dataset be constructed at scale while being resistant to pretraining approaches?

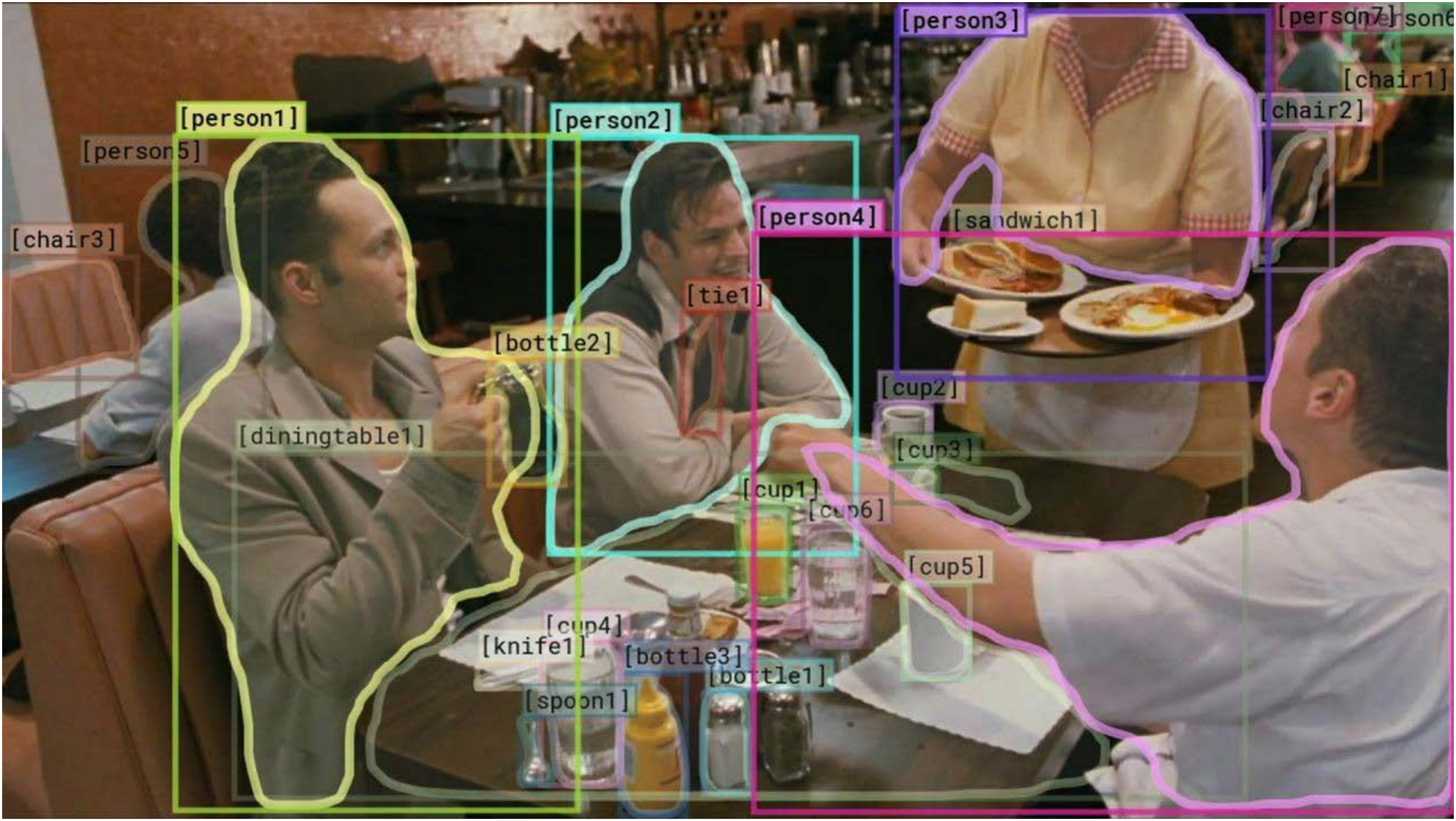
Can a 2 sentence NLI dataset be constructed at scale while being resistant to pretraining approaches?

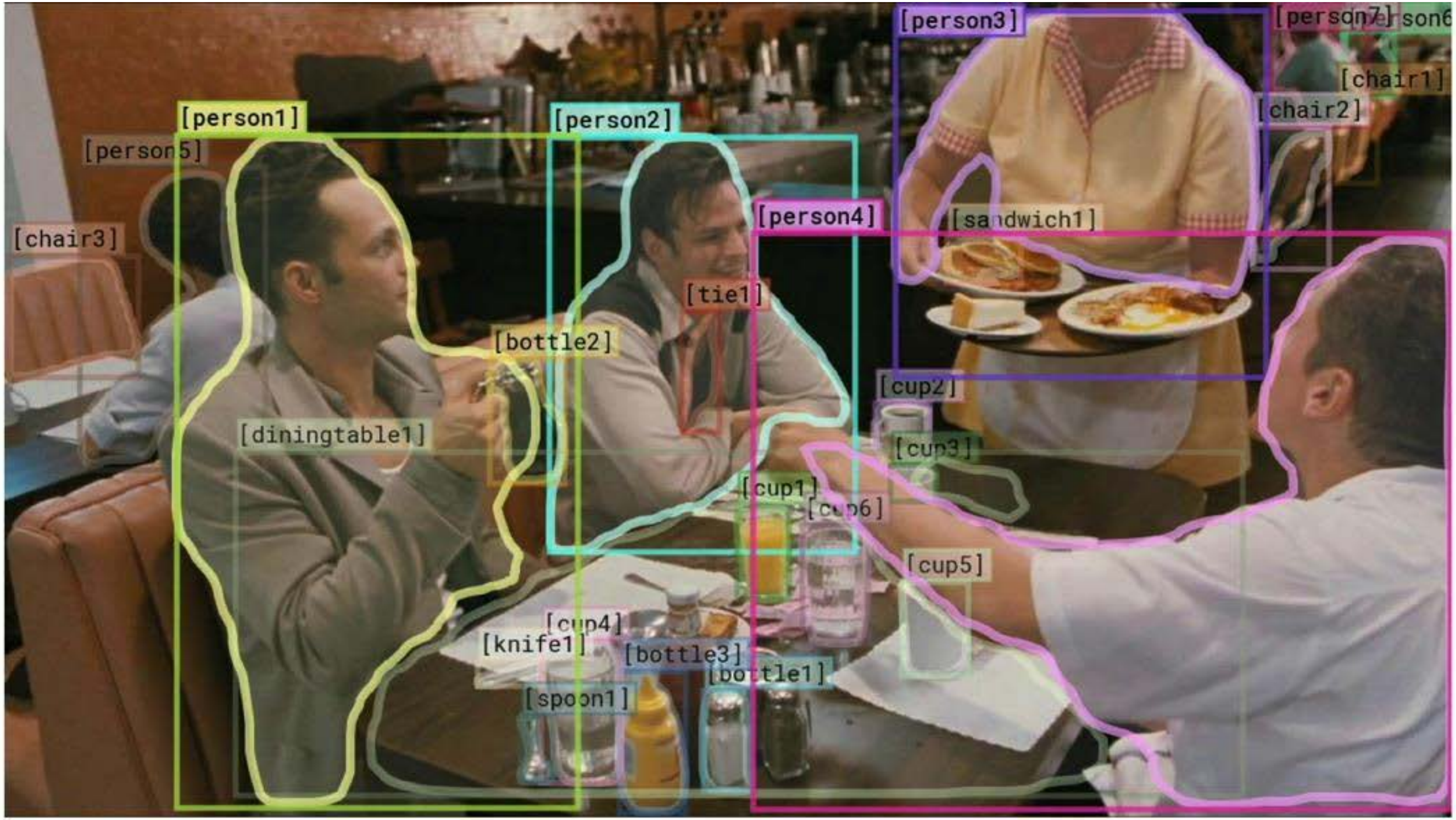
Actually, yes.


(work in progress)
BERT accuracy (4-way)
1.00
0.75
0.50
0.25
0.00
0
10
20
30
40
Adversarial Filtering iteration
SWAG1 (small LM)
SWAG2 (GPT)

Someone is pointing at someone else. The waiter…


MOVIECLIPS.COM


[person1]
[person2]
[person3]
[person4]
[person5]
[person7]
[chair1]
[chair2]
[chair3]
[sandwich1]
[tie1]
[bottle2]
[diningtable1]
[cup2]
[cup3]
[cup1]
[cup6]
[cup5]
[cup4]
[knife1]
[bottle3]
[bottle1]
[spoon1]



Why is [person4 ] pointing at [person1 ]?

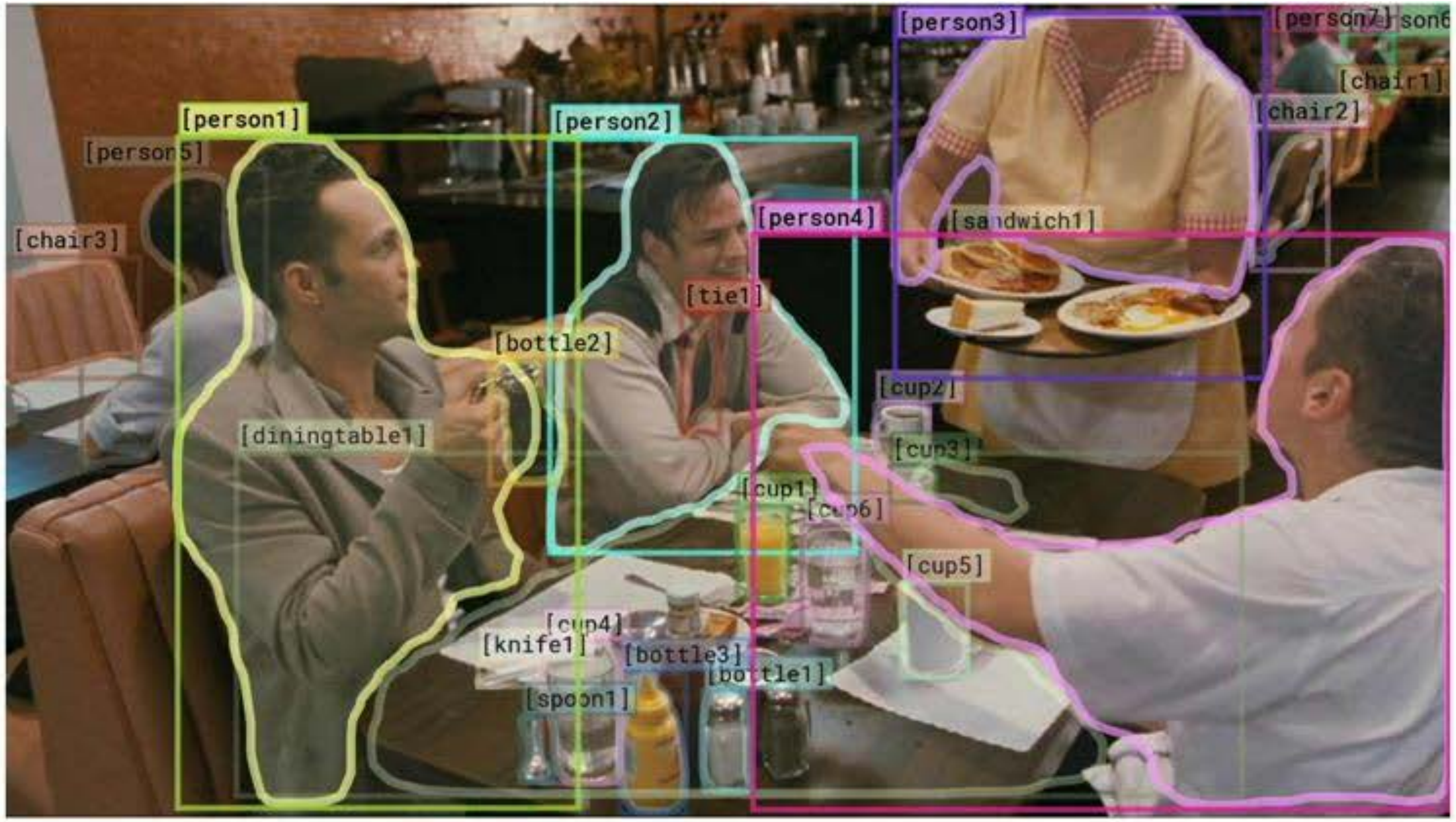
How can we get vision systems to learn this??

Why is [person4 ] pointing at [person1 ]?

From Recognition to Cognition: Visual Commonsense Reasoning (arxiv18)

VCR

VISUAL COMMONSENSE REASONING



Why is [person4] pointing at [person1]?

a) He is telling [person3] that [person1] ordered the pancakes.

b) He just told a joke.

c) He is feeling accusatory towards [person1].

d) He is giving [person1] directions.

Many AI systems perform well, but do so for ***questionable reasons***

Why is [person4] pointing at [person1]?

a) He is telling [person3] that [person1] ordered the pancakes.

b) He just told a joke.

c) He is feeling accusatory towards [person1].

d) He is giving [person1] directions.

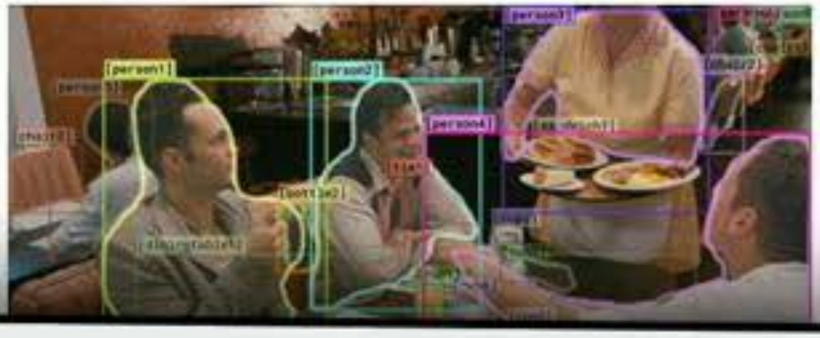
I chose **a)** because...

a) [person1] has the pancakes in front of him.

b) [person4] is taking everyone's order and asked for clarification.

c) [person3] is looking at the pancakes and both she and [person2] are smiling slightly.

d) [person3] is delivering food to the table, and she might not know whose order is whose.



How can we collect challenging commonsense inferences?

Why is ?

a) **cakes.**

b)

c) He is feeling accusatory towards [person1].

d) He is giving

How do we get the wrong answers?

How do we model this?

pancakes in

everyone's n.

at the pa

are smiling slightly.

d) [person3] is delivering food to the table, and she **ot know whose order is whose.**

Our contributions

- New task: Visual Commonsense Reasoning 

VCR
VISUAL COMMONSENSE REASONING

Our contributions

- New task: Visual Commonsense Reasoning

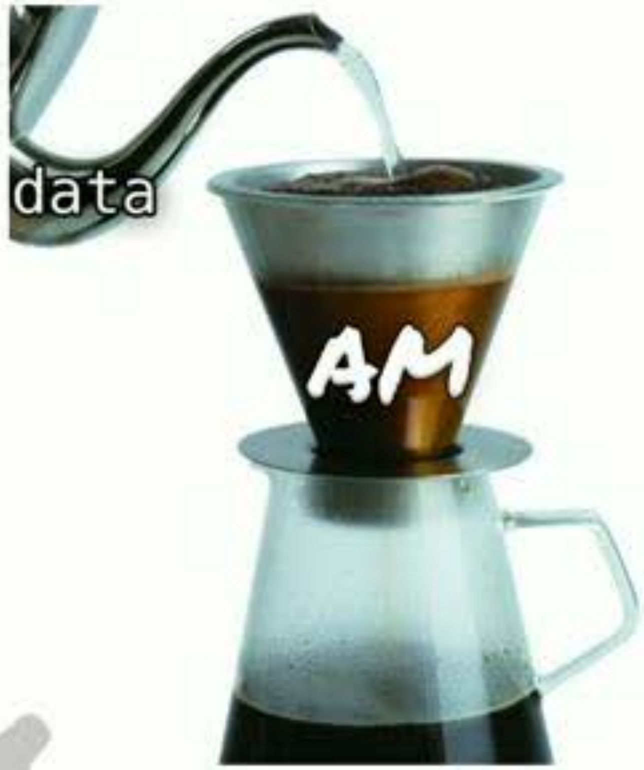
- Building VCR, feat. Adversarial Matching

- Recognition to Cognition Networks

Our contributions

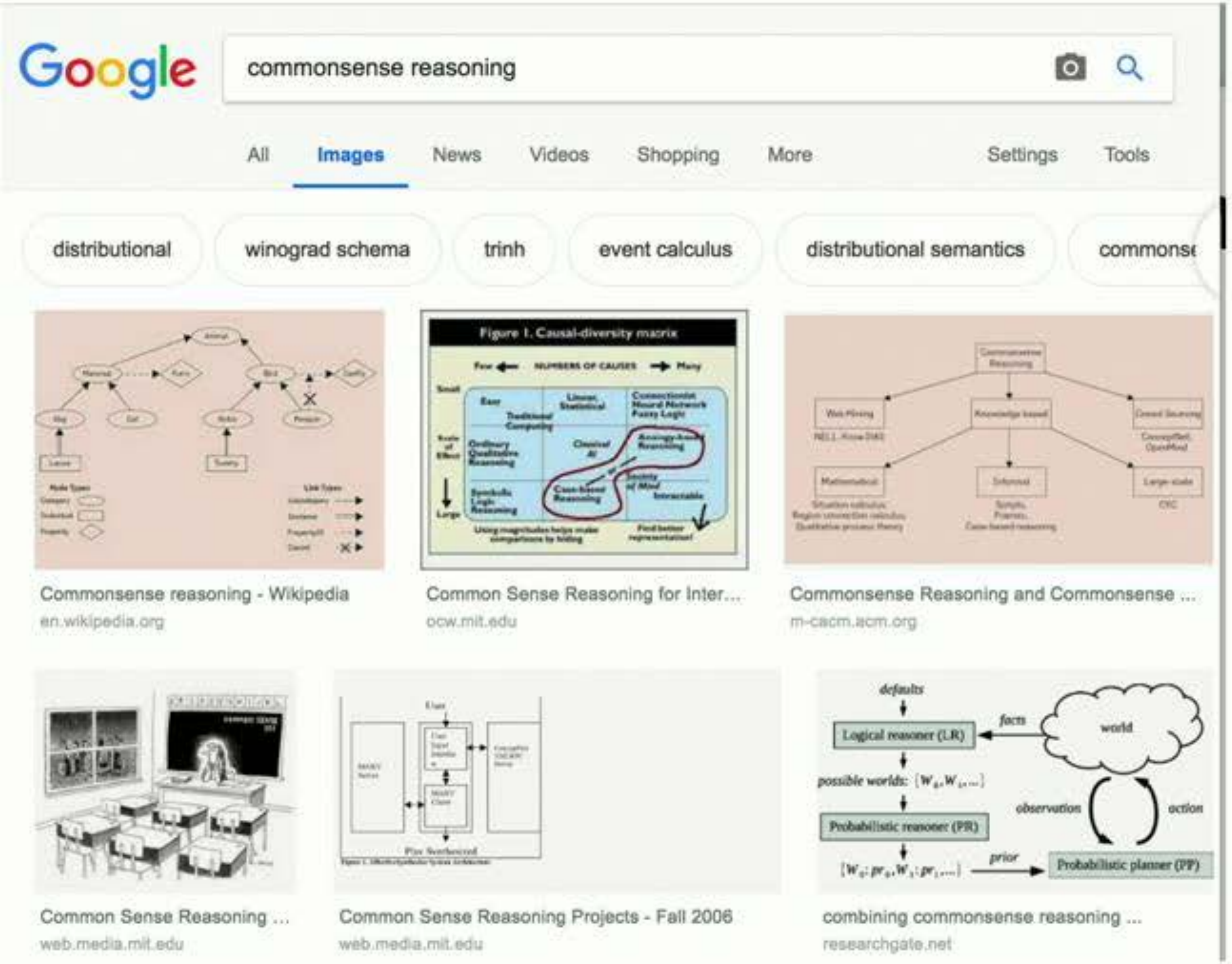
- New task: Visual Commonsense Reasoning

- Building VCR, feat. Adversarial Matching



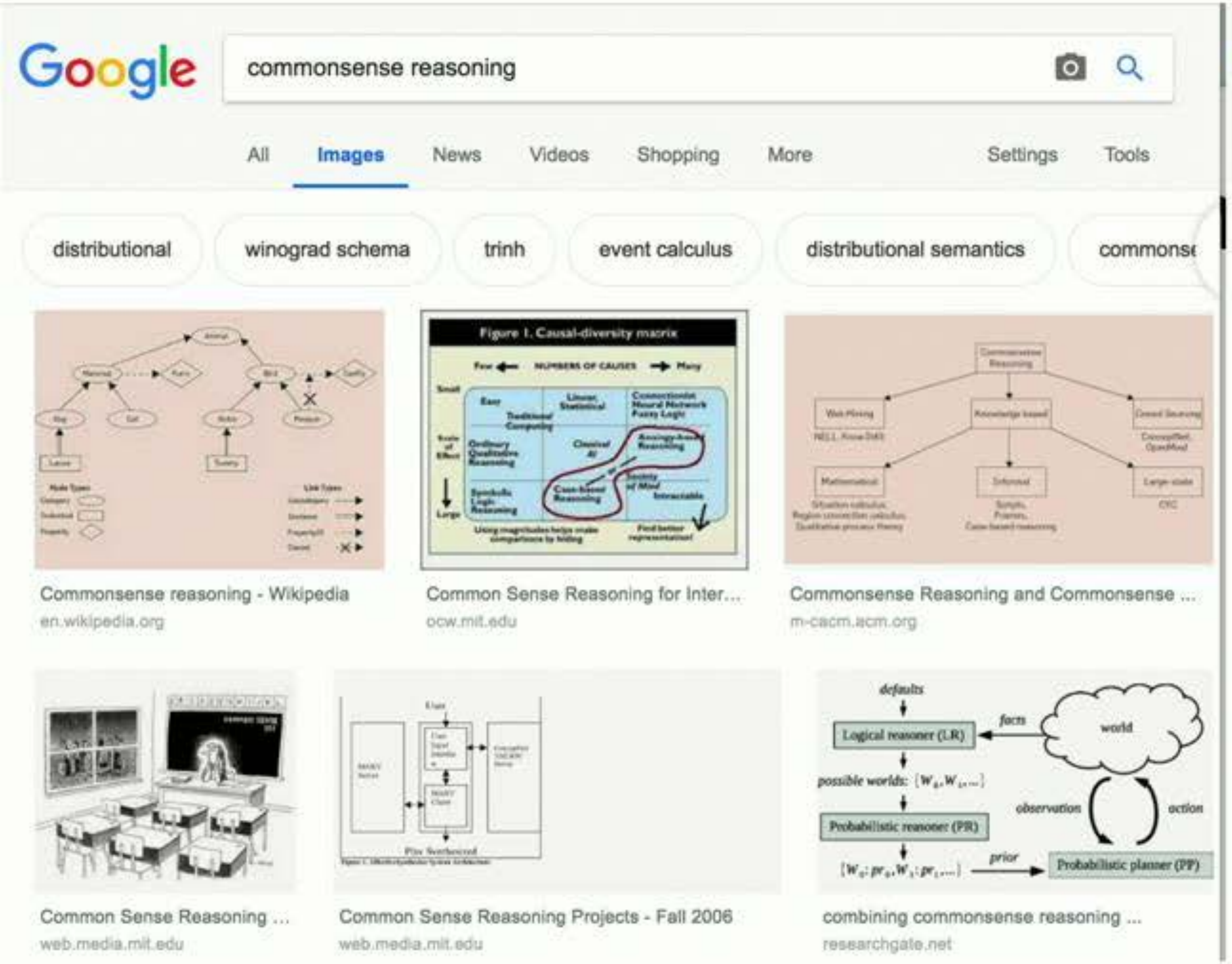
- Recognition to Cognition Networks

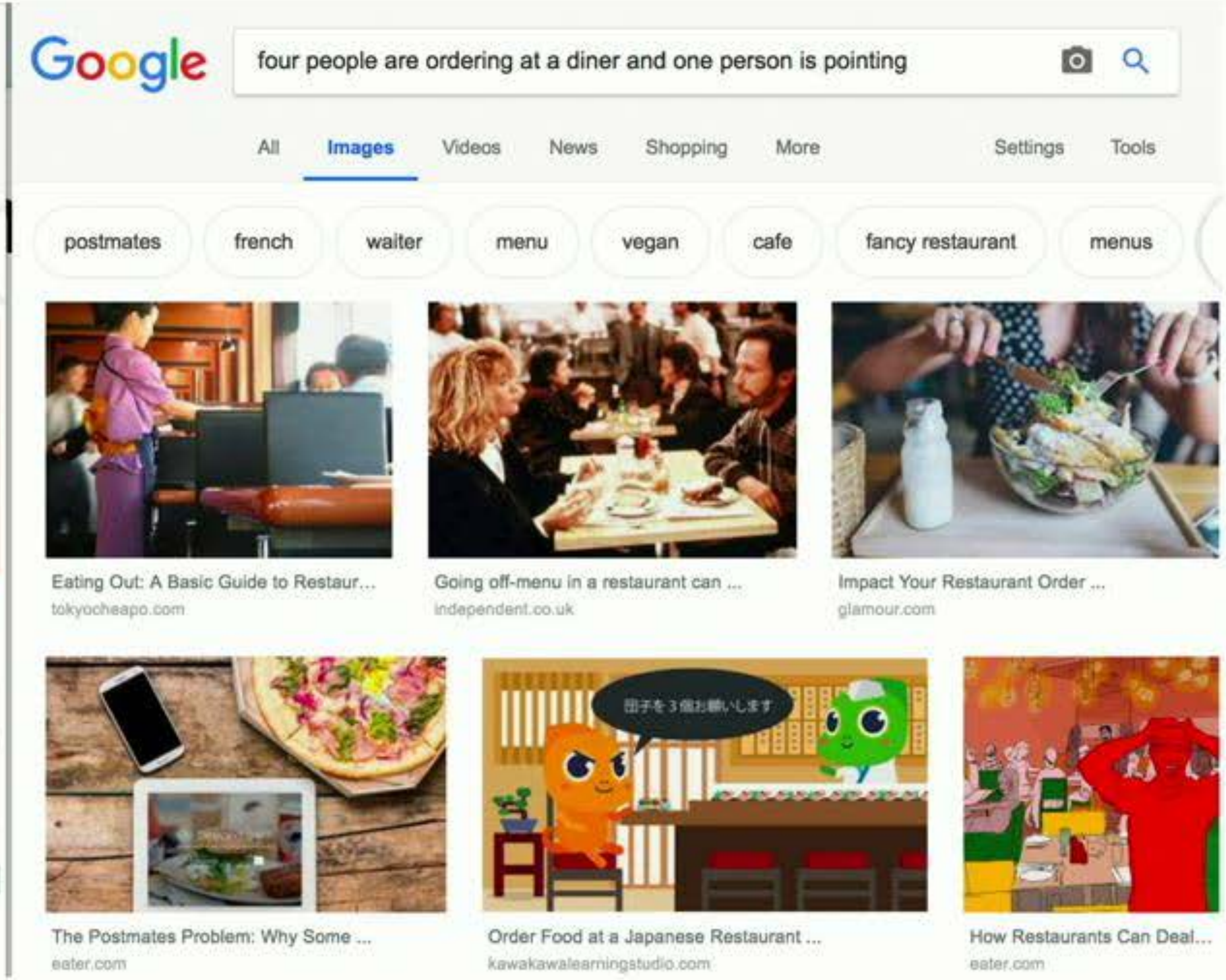
Collecting commonsense inferences



fouhey et al, "from lifestyle vlogs to everyday interactions." cvpr2018

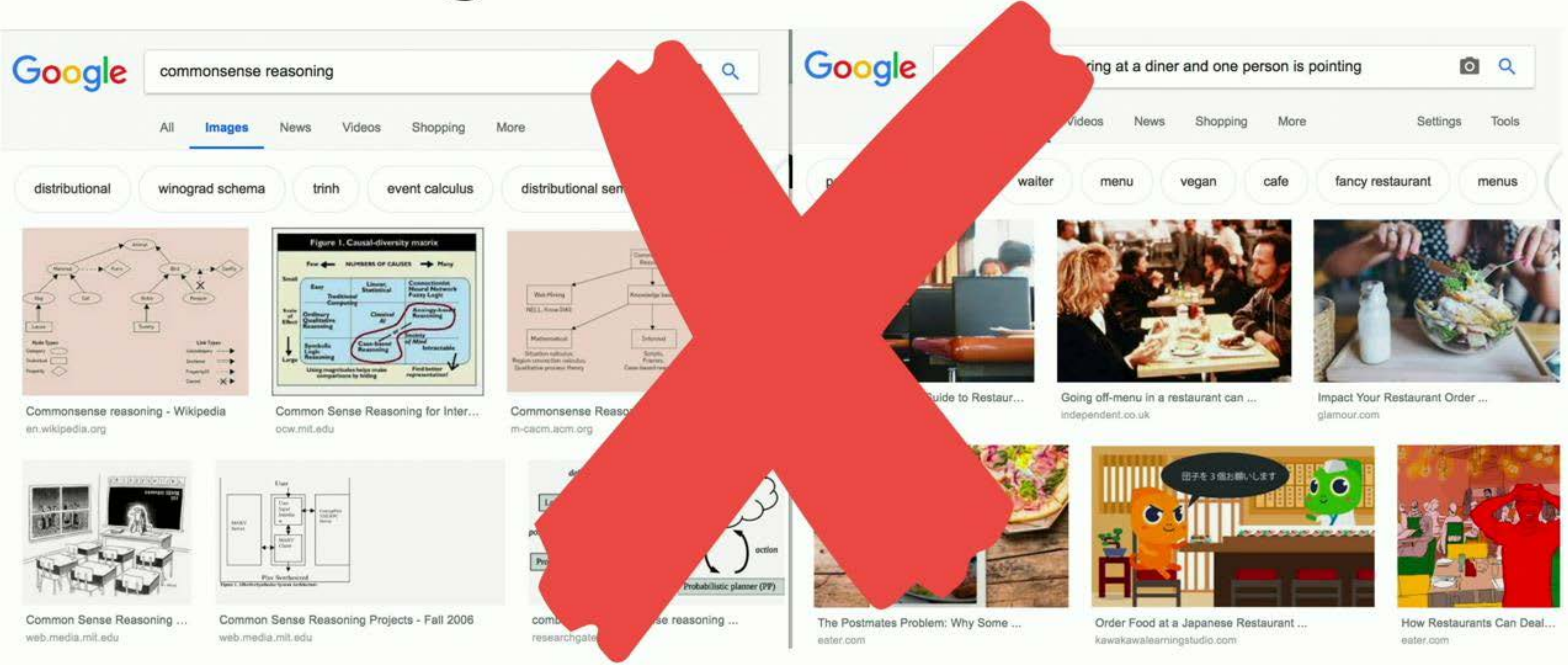
Collecting commonsense inferences





fouhey et al, “from lifestyle vlogs to everyday interactions.” cvpr2018

Collecting commonsense inferences



fouhey et al, “from lifestyle vlogs to everyday interactions.” cvpr2018

Collecting commonsense inferences

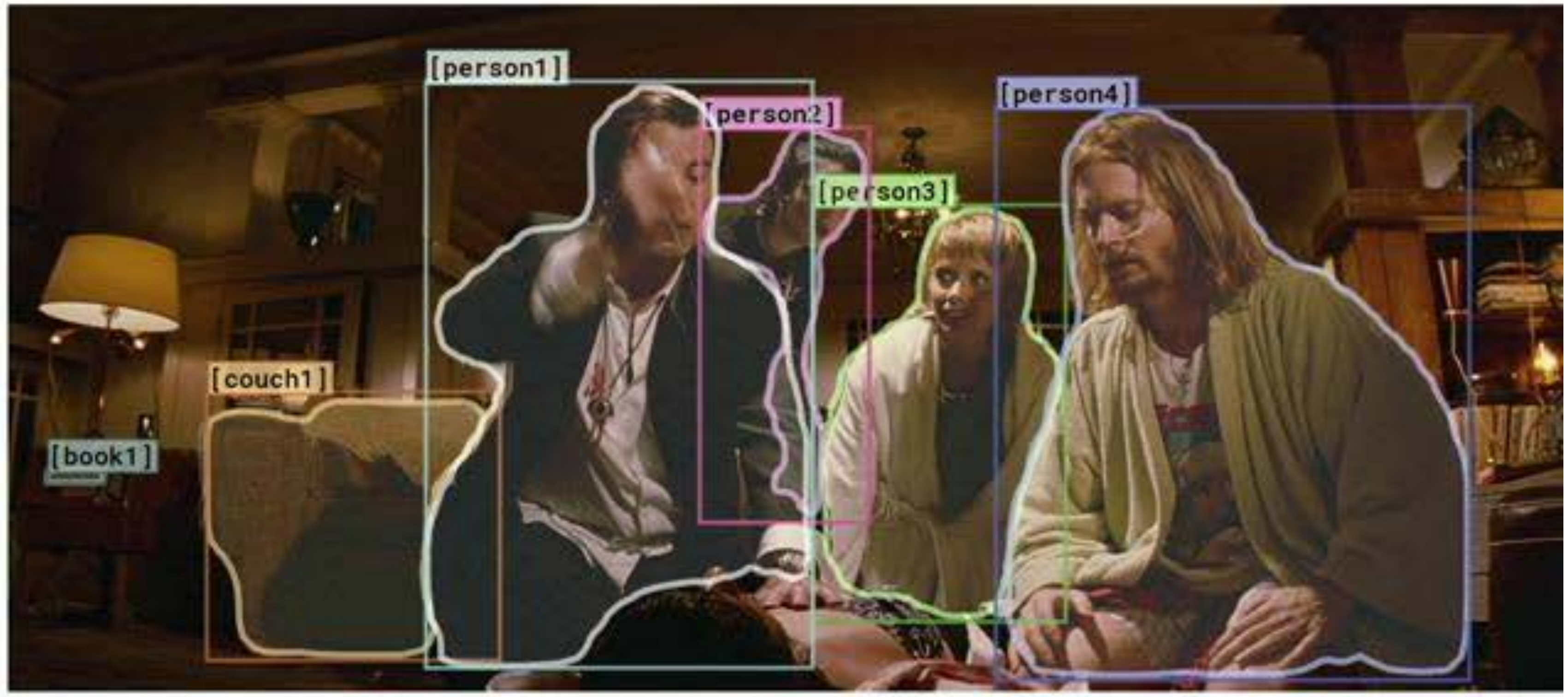
LSMDC

MOVIECLIPS
we love movies

Collecting commonsense inferences

LSMDC MOVIECLIPS we love movies

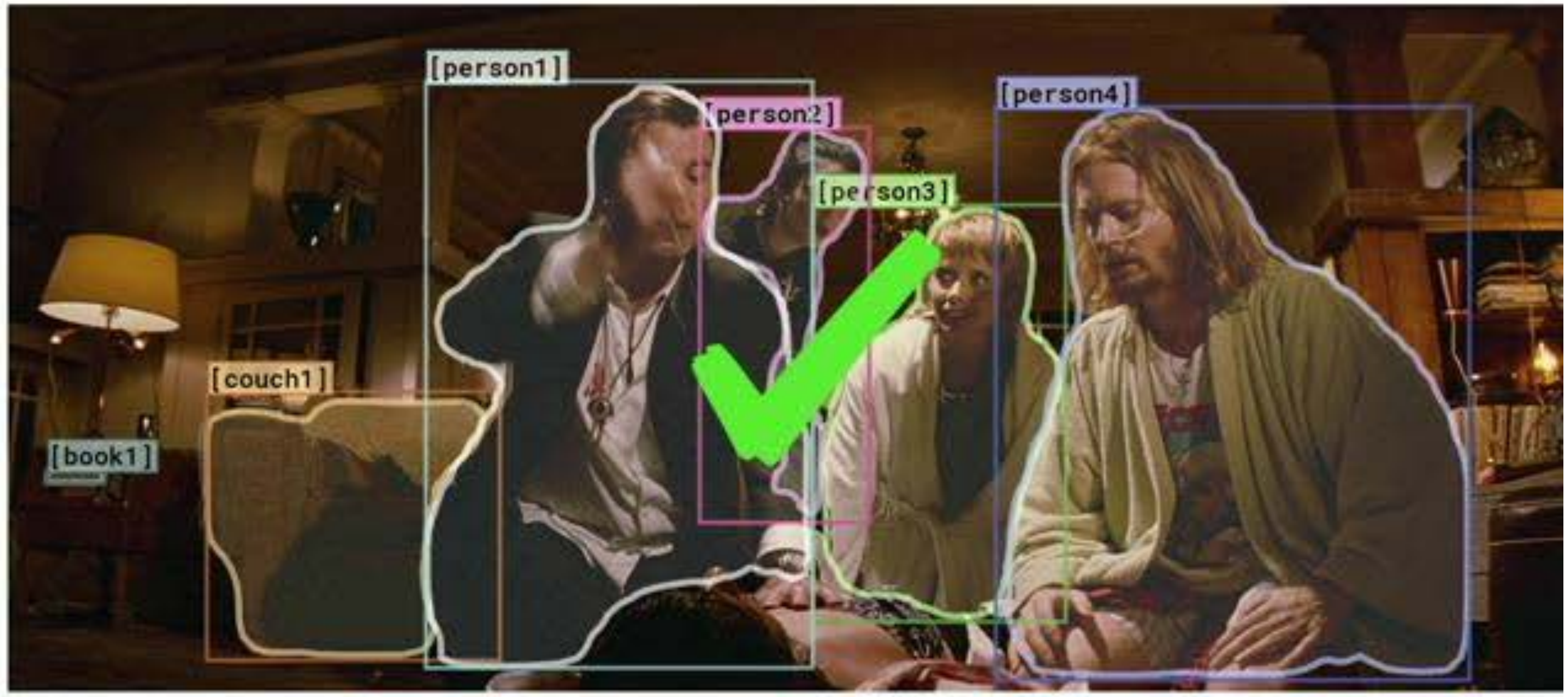
Collecting commonsense inferences



LSMDC

MOVIECLIPS
we love movies

Collecting commonsense inferences



LSMDC

MOVIECLIPS
we love movies

Collecting commonsense inferences

Examples (see rowanzellers.com/halloffame for more) (expand/collapse)

Past caption: The driver looks out and sees the Ducati's front wheel waggling about in midair.

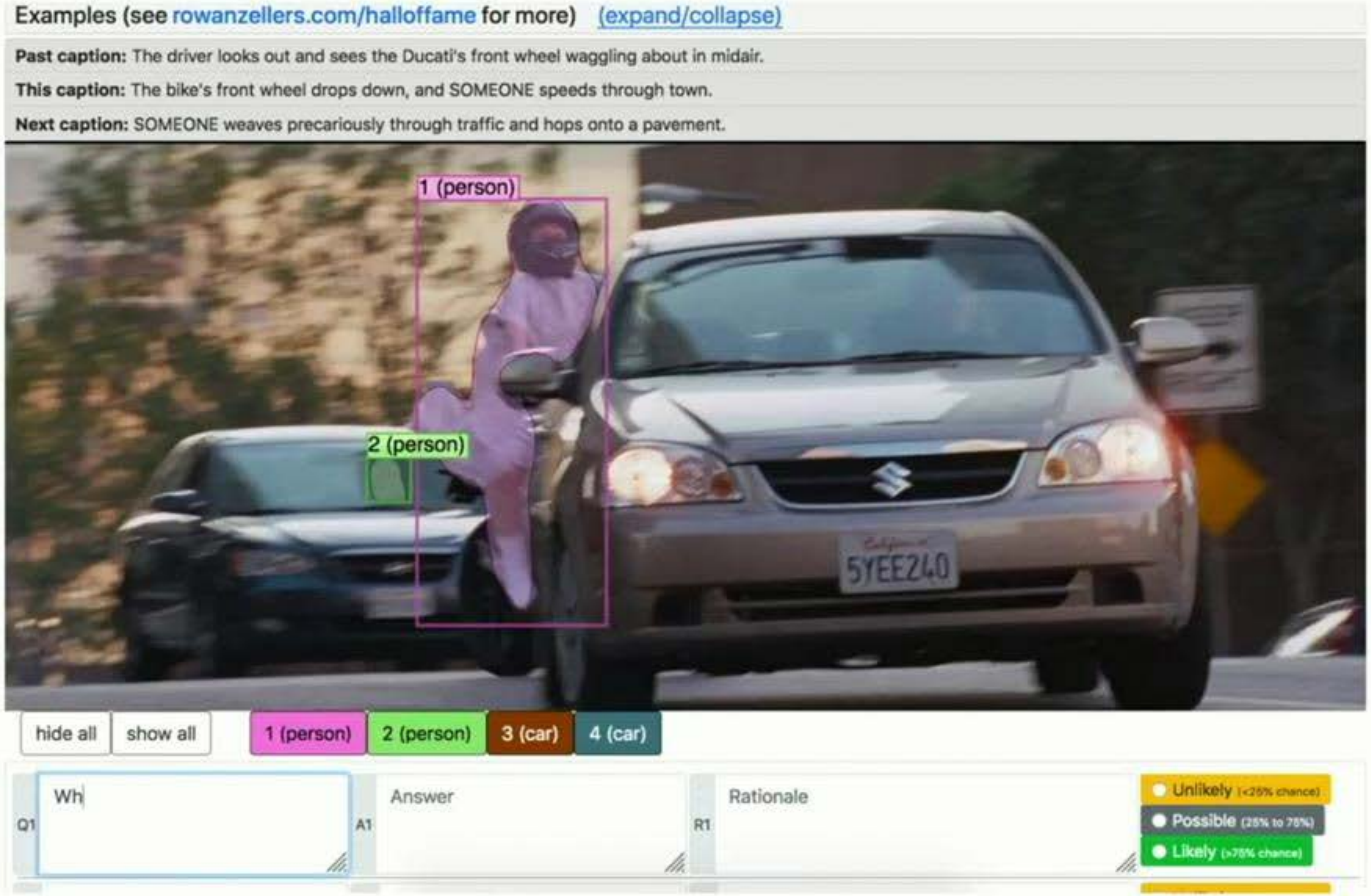
This caption: The bike's front wheel drops down, and SOMEONE speeds through town.

Next caption: SOMEONE weaves precariously through traffic and hops onto a pavement.

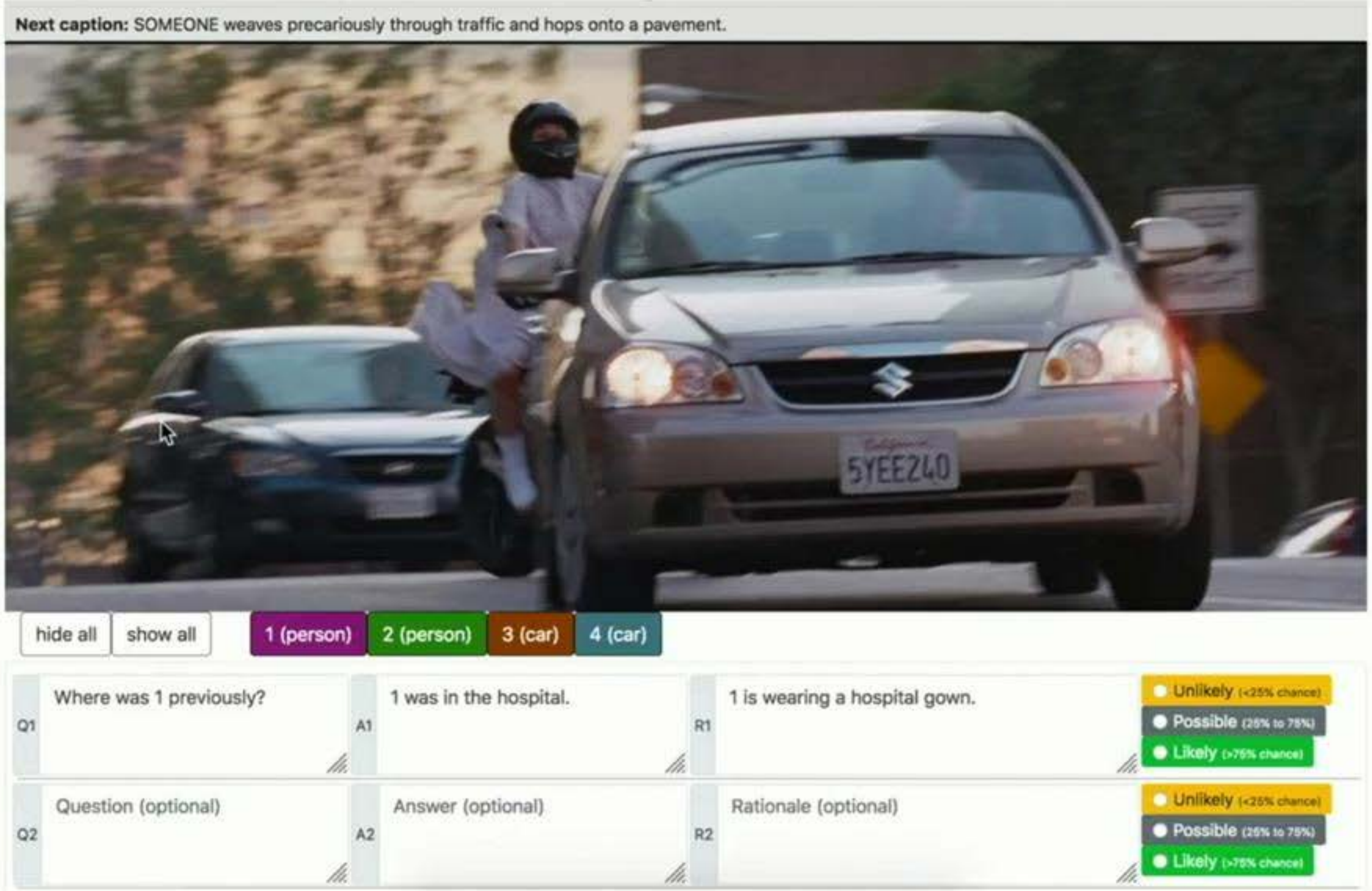
hide all | show all | 1 (person) | 2 (person) | 3 (car) | 4 (car)

Q1	A1	R1	
Question	Answer	Rationale	Unlikely (<25% chance) Possible (25% to 75%) Likely (>75% chance)

Collecting commonsense inferences



Collecting commonsense inferences



How do we get the wrong answers, avoiding annotation artifacts?

what color is the shirt?

what color is the shirt?

teal

blue

violet

magenta

Adversarial Matching

Wrong answers must have be

relevant to the question yet different from the correct answer.

Adversarial Matching

Wrong answers must have be

relevant to the question yet different from the correct answer.

Q,A' ***Question Relevance***

A,A' ***Entailment***

Adversarial Matching

Wrong answers must have be

relevant to the question yet different from the correct answer.

Q,A' **Question Relevance**

A,A' **Entailment**

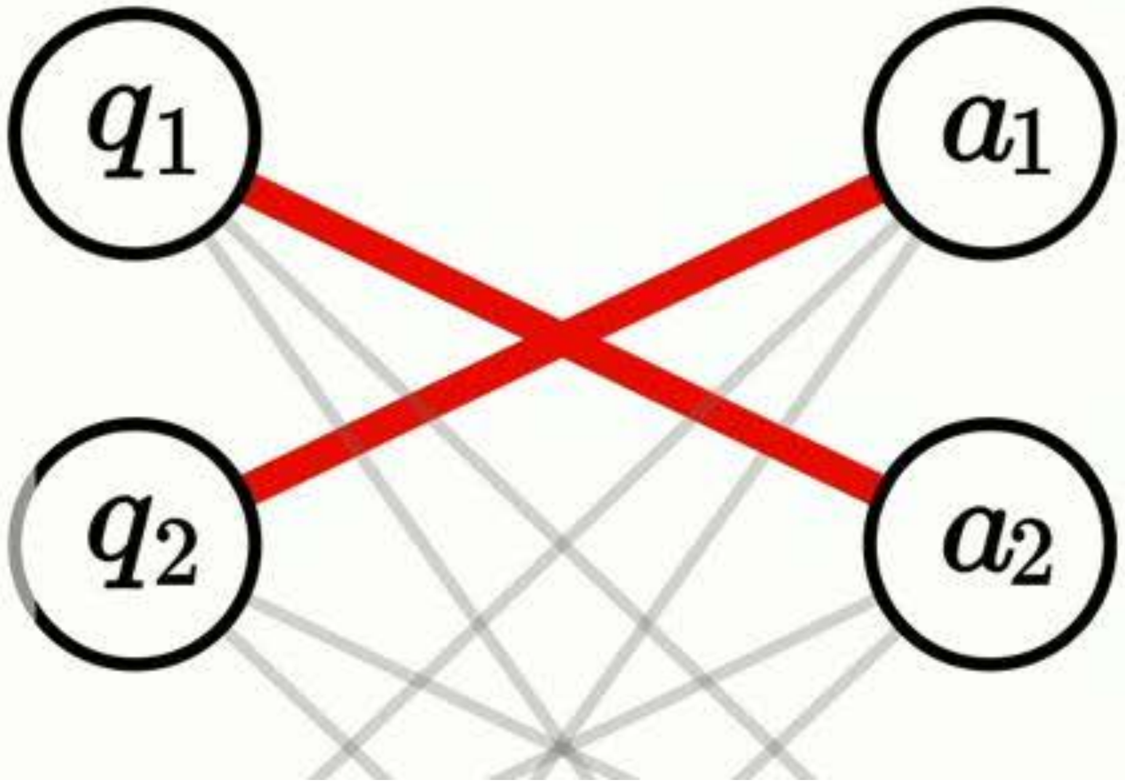
We'll use these two metrics to **recycle right answers** to other questions, using a minimum weight bipartite matching.

Adversarial Matching

We'll use these two metrics to **<u>recycle right answers</u>** to other questions, using a minimum weight bipartite matching.

Why are [person1] and [person3] holding their foreheads together?

Why do [person1] and [person3] have their hands clasped?



They are about to kiss.

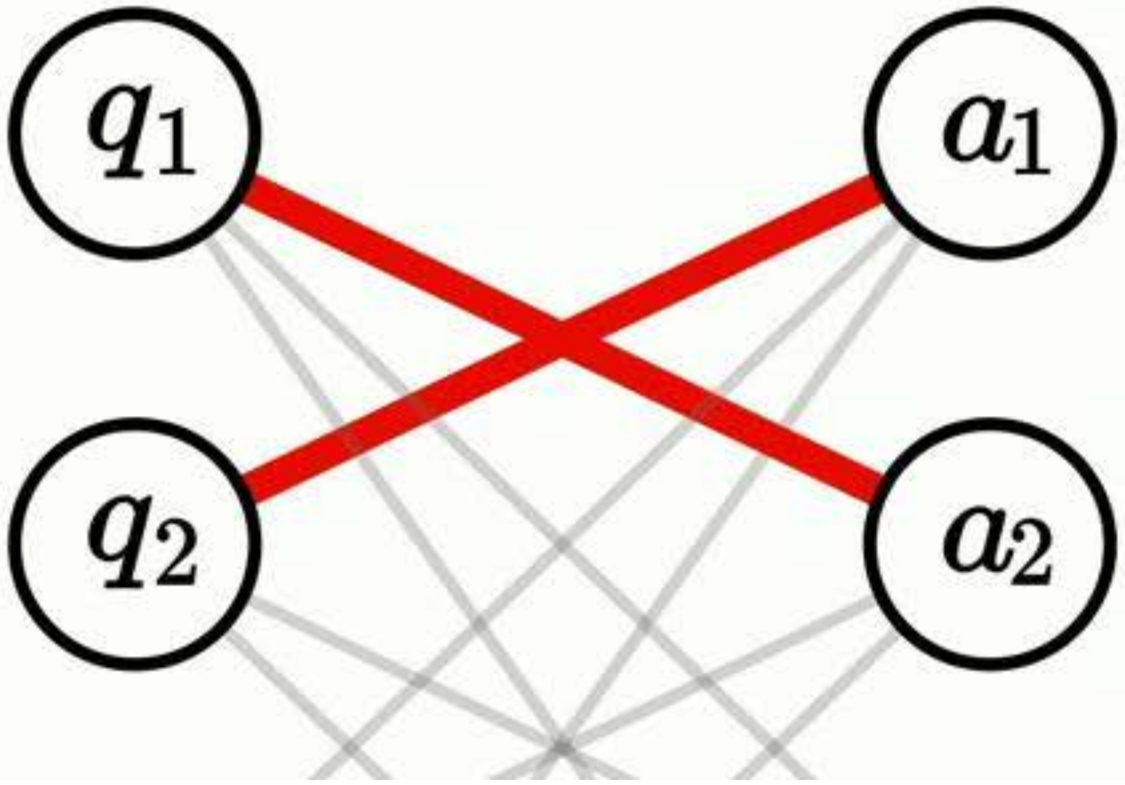
[person1] and [person3] are praying.

Adversarial Matching

We'll use these two metrics to **recycle right answers** to other questions, using a minimum weight bipartite matching.

Why are [person1] and [person3] holding their foreheads together?

Why do [person1] and [person3] have their hands clasped?



They are about to kiss.

[person1] and [person3] are praying.

Adversarial Matching

Wrong answers must have be

relevant to the question yet different from the correct answer.

Q,A' ***Question Relevance***

A,A' ***Entailment***

Adversarial Matching

Wrong answers must have be

relevant to the question   yet   different from the correct answer.

Q, A' **Question Relevance**

A, A' **Entailment**

Question relevance dominates - hard for machines

Entailment penalty dominates - easy for humans

Adversarial Matching

Wrong answers must have be

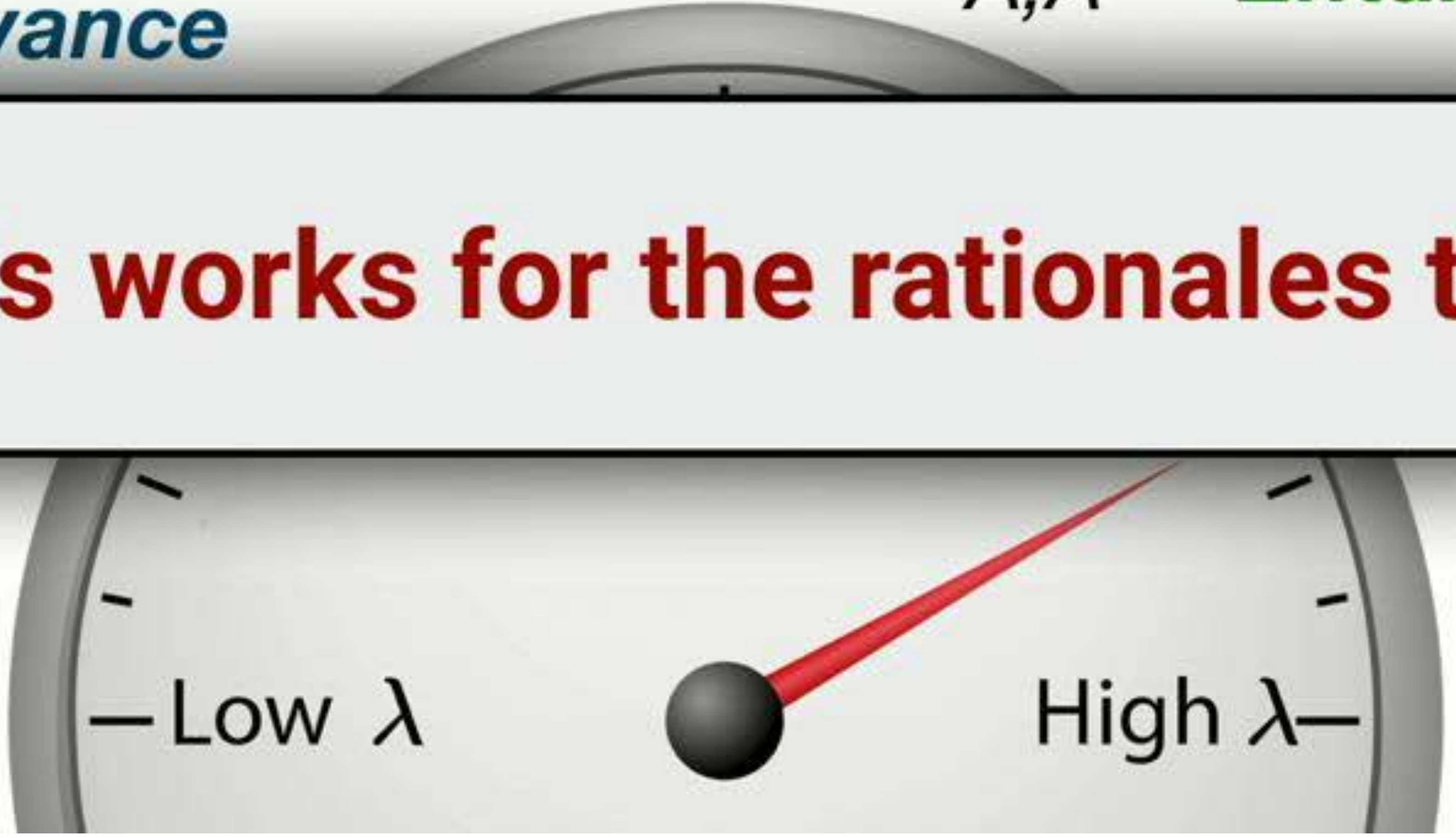
relevant to the question yet different from the correct answer.

Q,A' ***Question Relevance***

A,A' ***Entailment***

This works for the rationales too!

Question relevance dominates - hard for machines



Entailment penalty dominates - easy for humans

What about the people tags (`[person5]`)?

Adversarial Matching

Wrong answers must have be

relevant to the question yet different from the correct answer.

Q,A' **Question Relevance**

A,A' **Entailment**

Adversarial Matching

Wrong answers must have be

relevant to the question yet different from the correct answer.

Q,A' **Question Relevance**

A,A' **Entailment**

Question relevance dominates - hard for machines

Entailment penalty dominates - easy for humans

Adversarial Matching

Wrong answers must have be

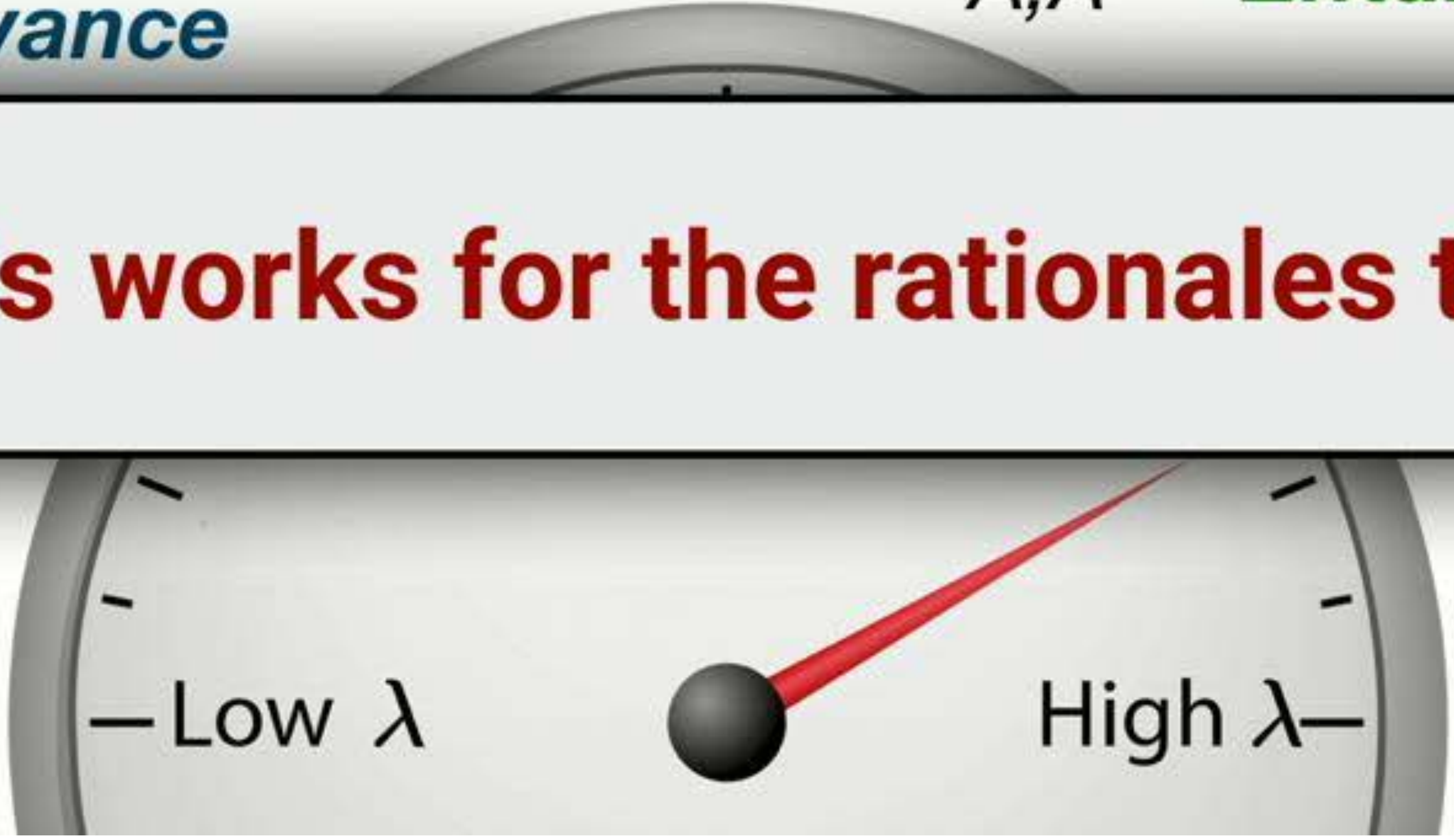
relevant to the question yet different from the correct answer.

Q,A' **Question Relevance**

A,A' **Entailment**

This works for the rationales too!

Question relevance dominates - hard for machines



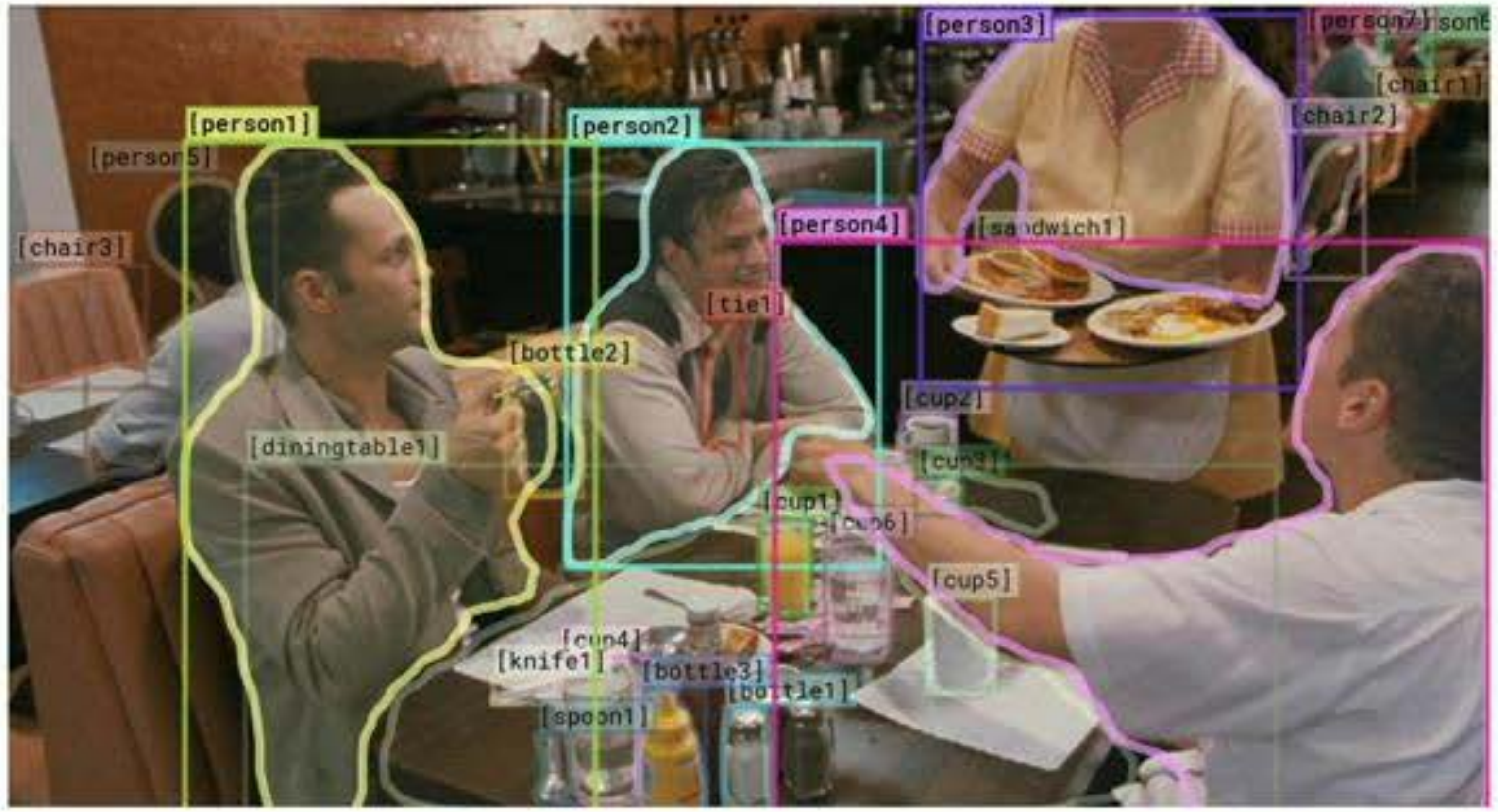
Entailment penalty dominates - easy for humans

What about the people tags (`[person5]`)?

We’ll randomly modify the detection tags in candidate answer to better match the new question/image.

What about the people tags (`[person5]`)?

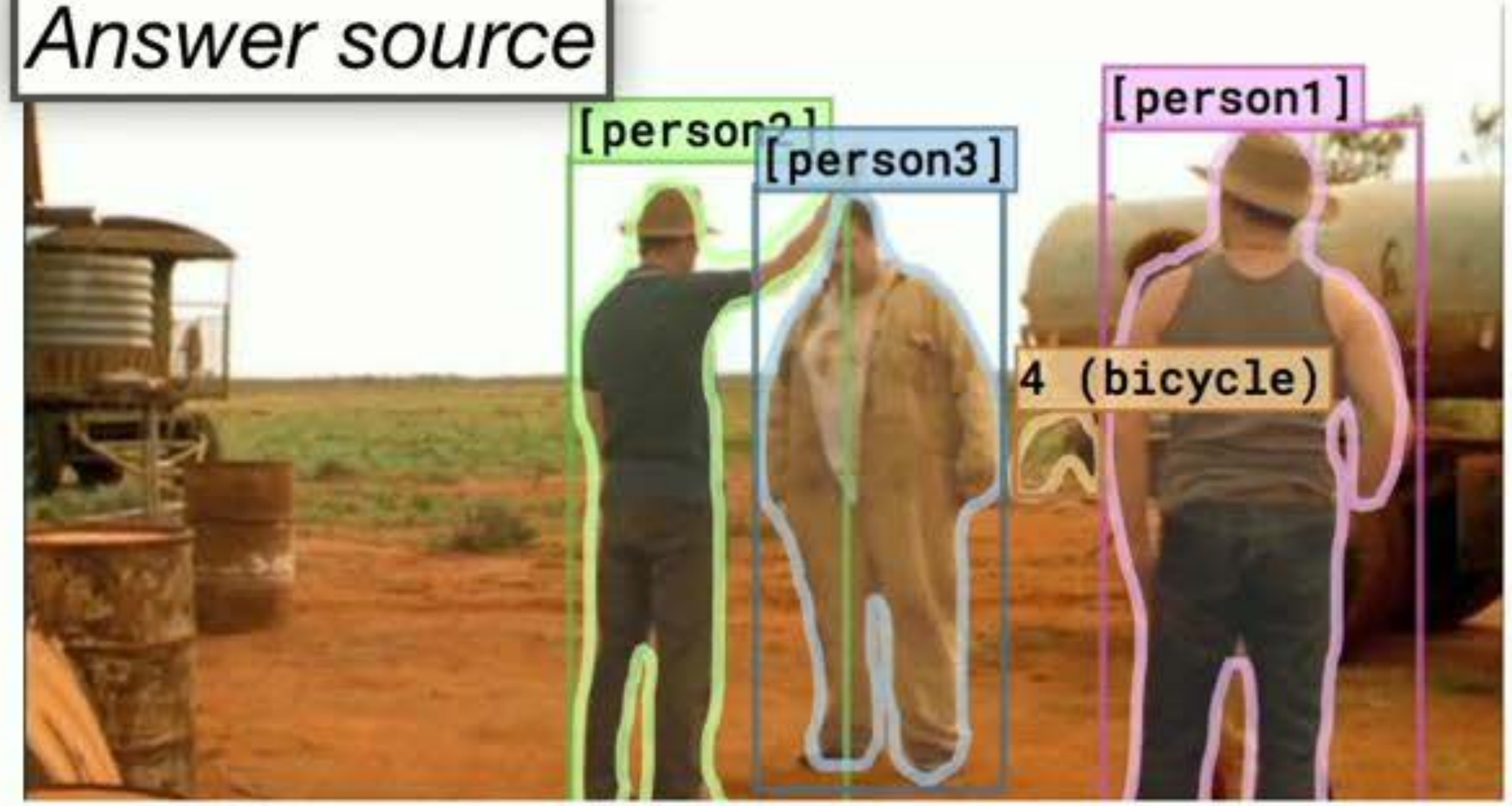
We'll randomly modify the detection tags in candidate answer to better match the new question/image.

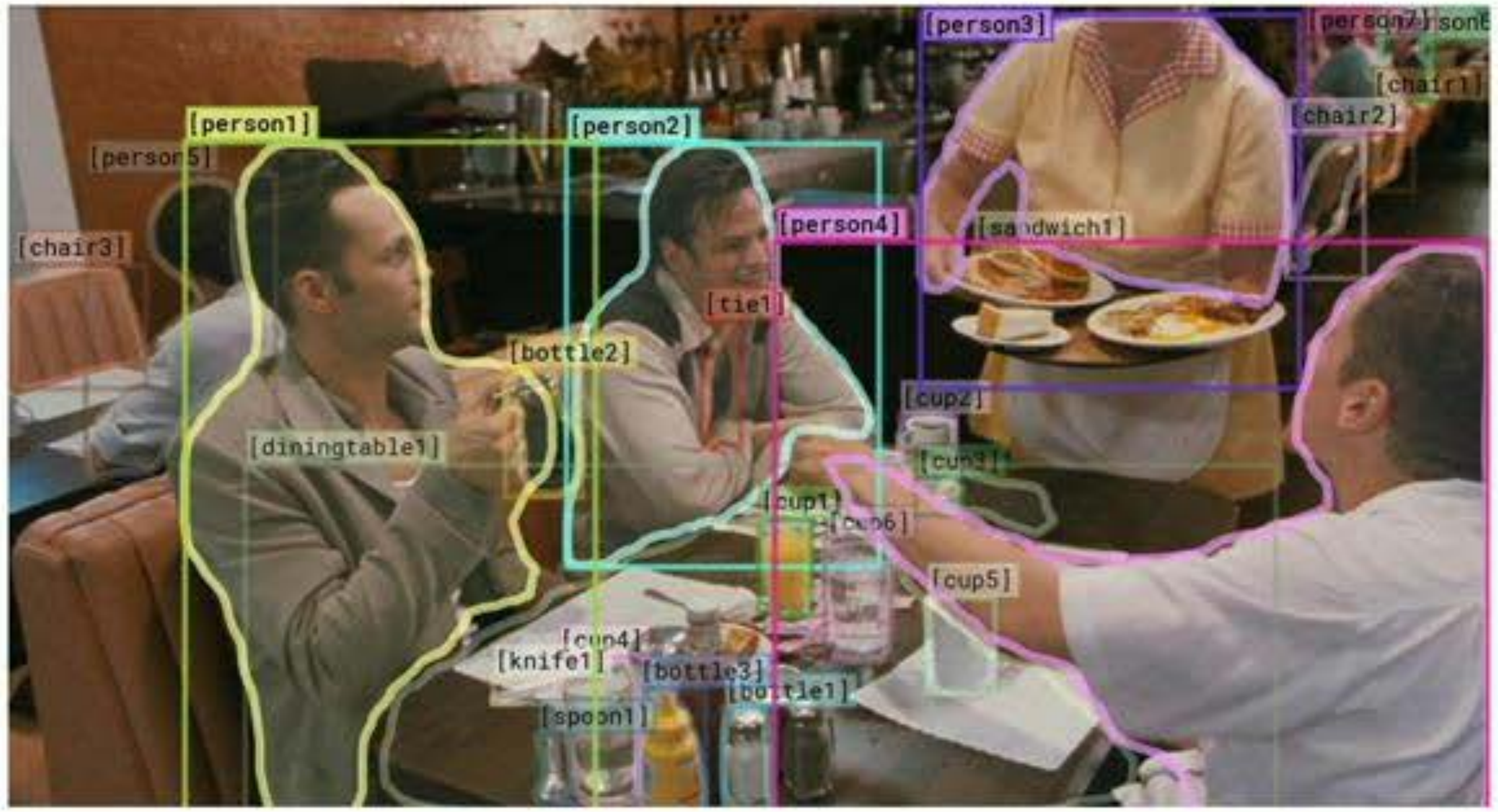


What about the people tags ([person5])?

We'll randomly modify the detection tags in candidate answer to better match the new question/image.

Answer source



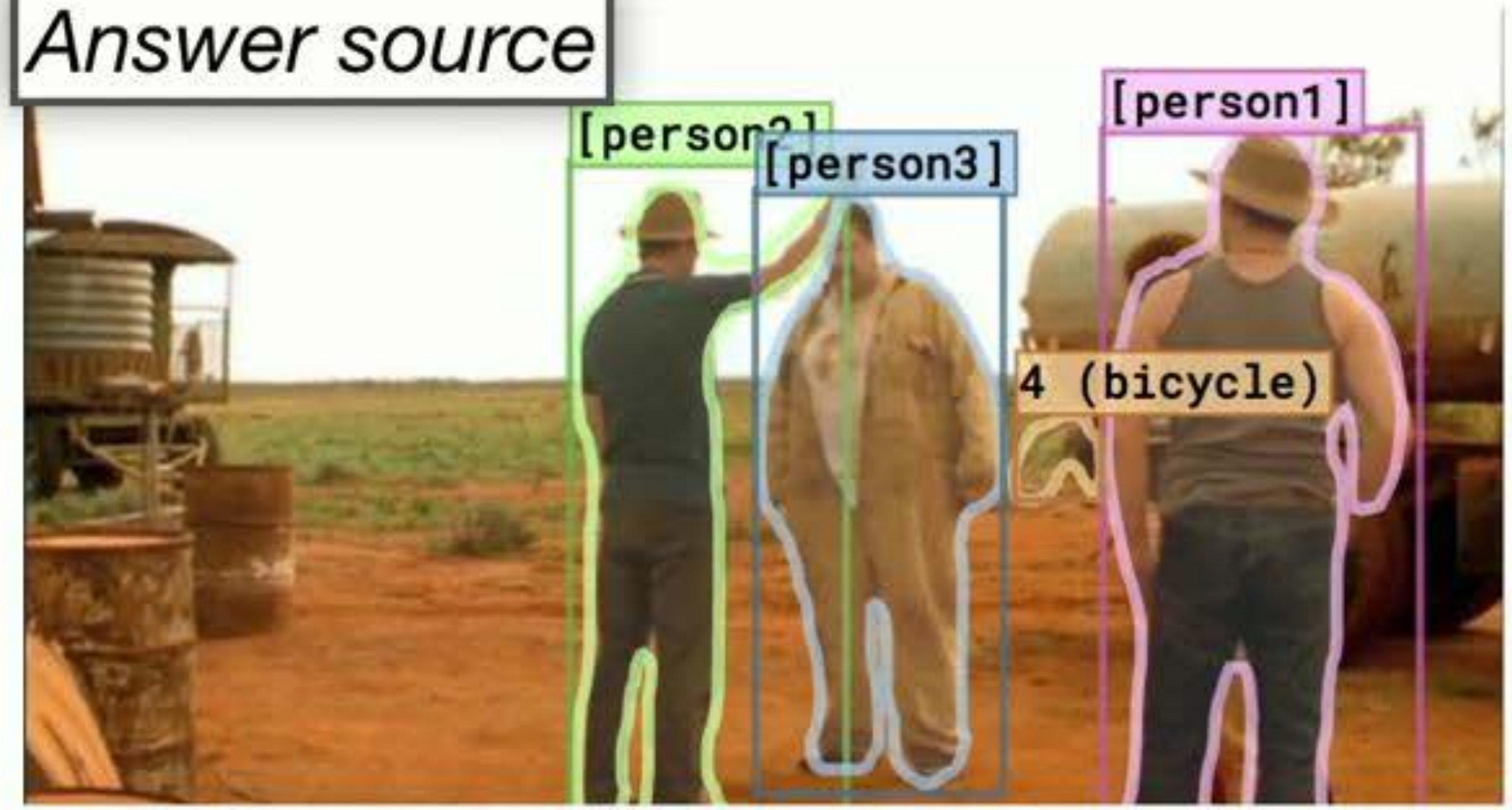


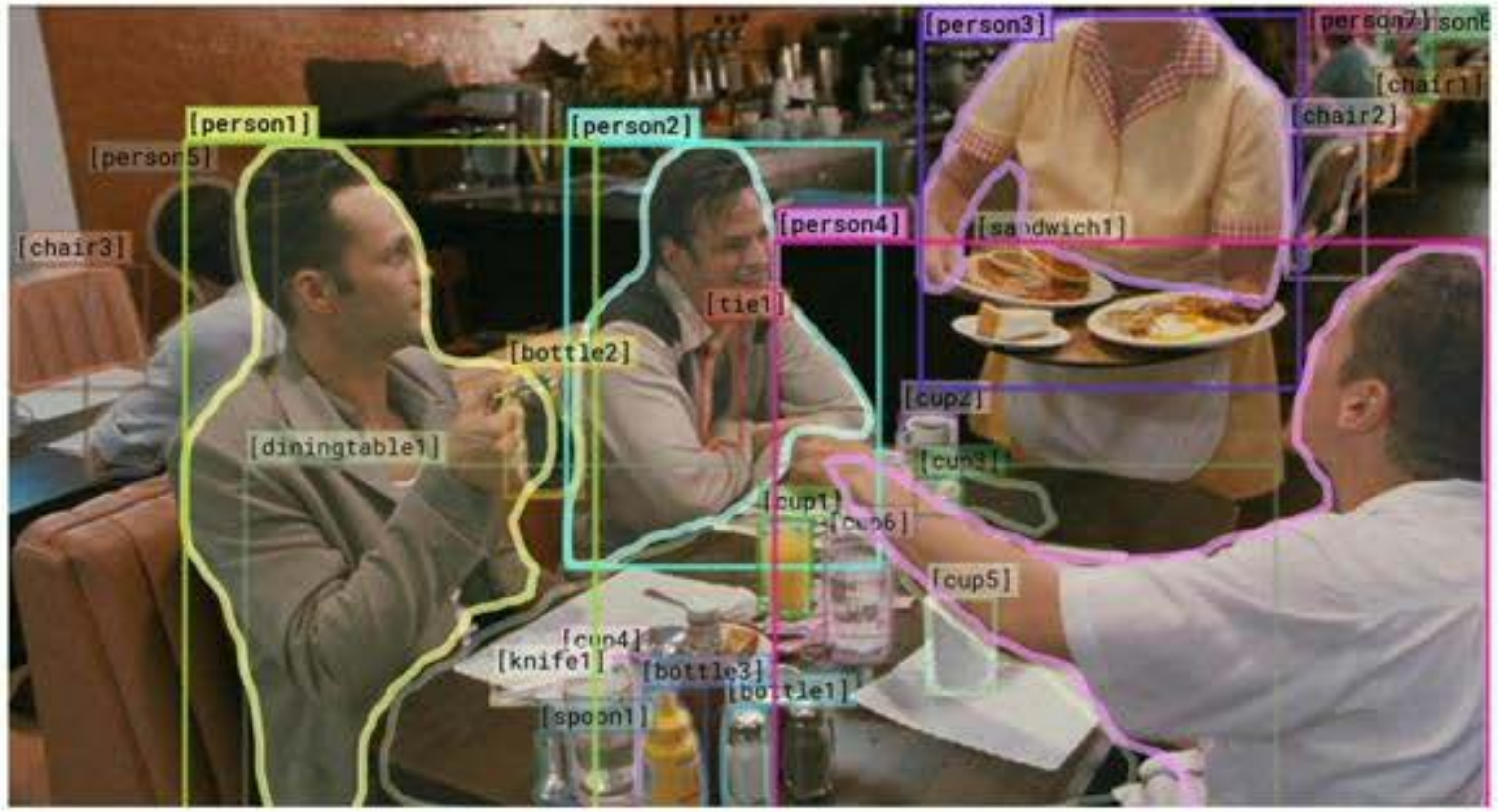
He is giving **[person3**  **]** directions.

What about the people tags ([person5])?

We'll randomly modify the detection tags in candidate answer to better match the new question/image.

Answer source





He is giving **[person3]** directions.

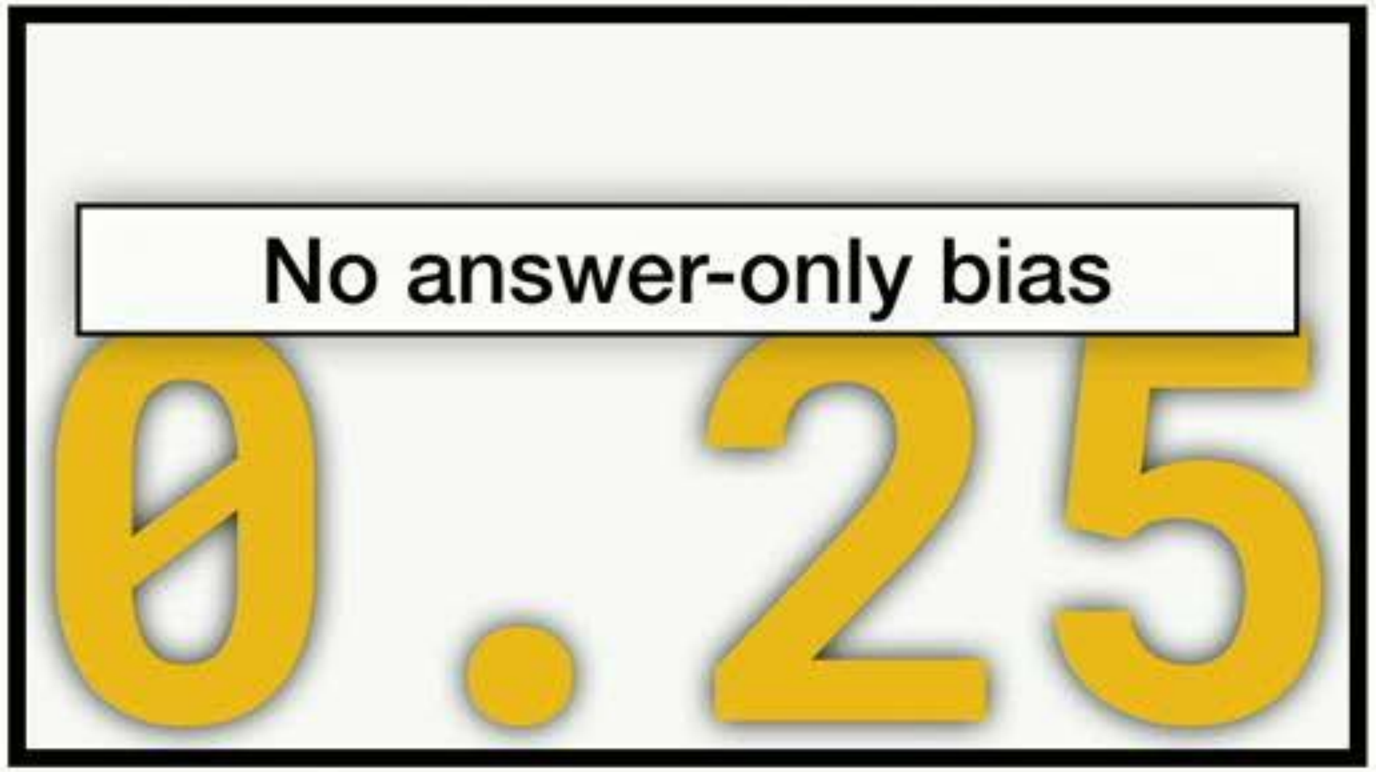
He is giving **[person1]** directions.

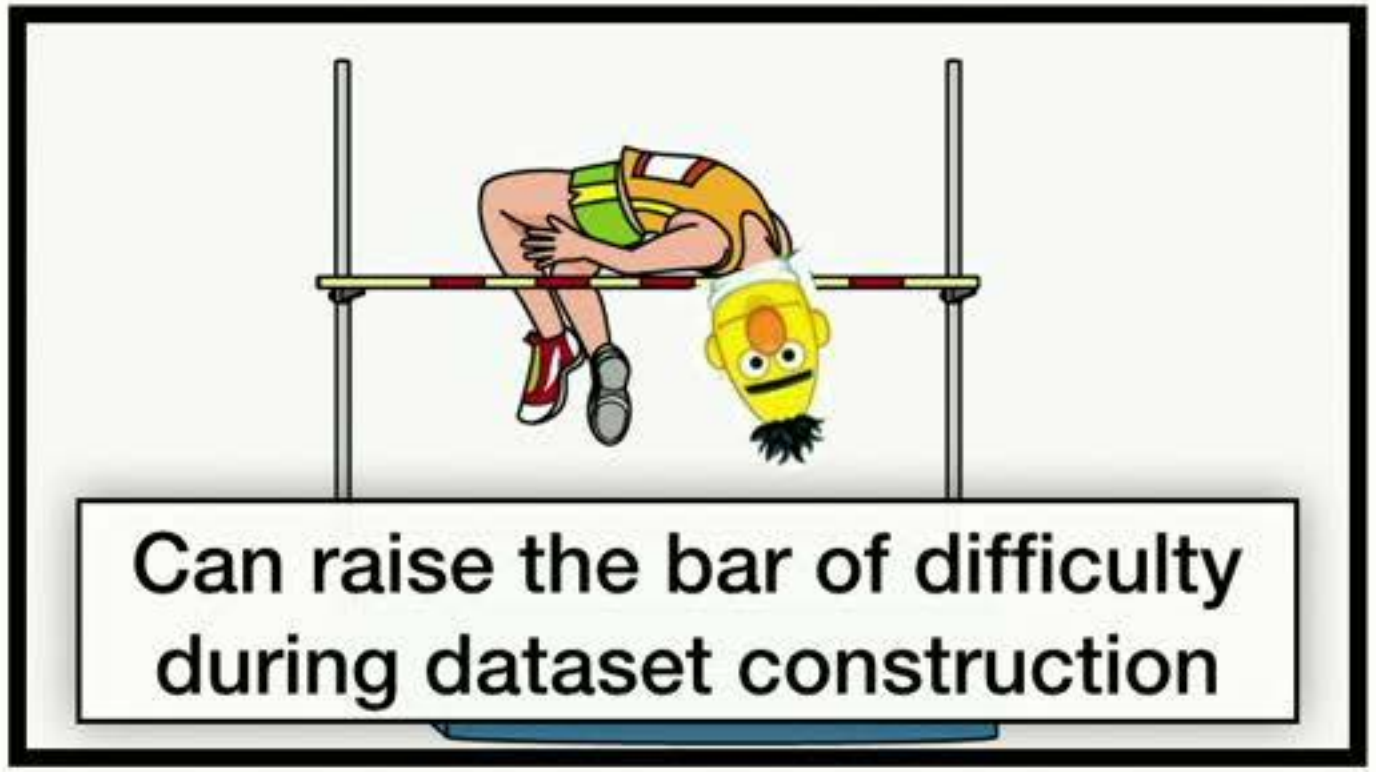
Unique Contributions of Adversarial Matching

No answer-only bias

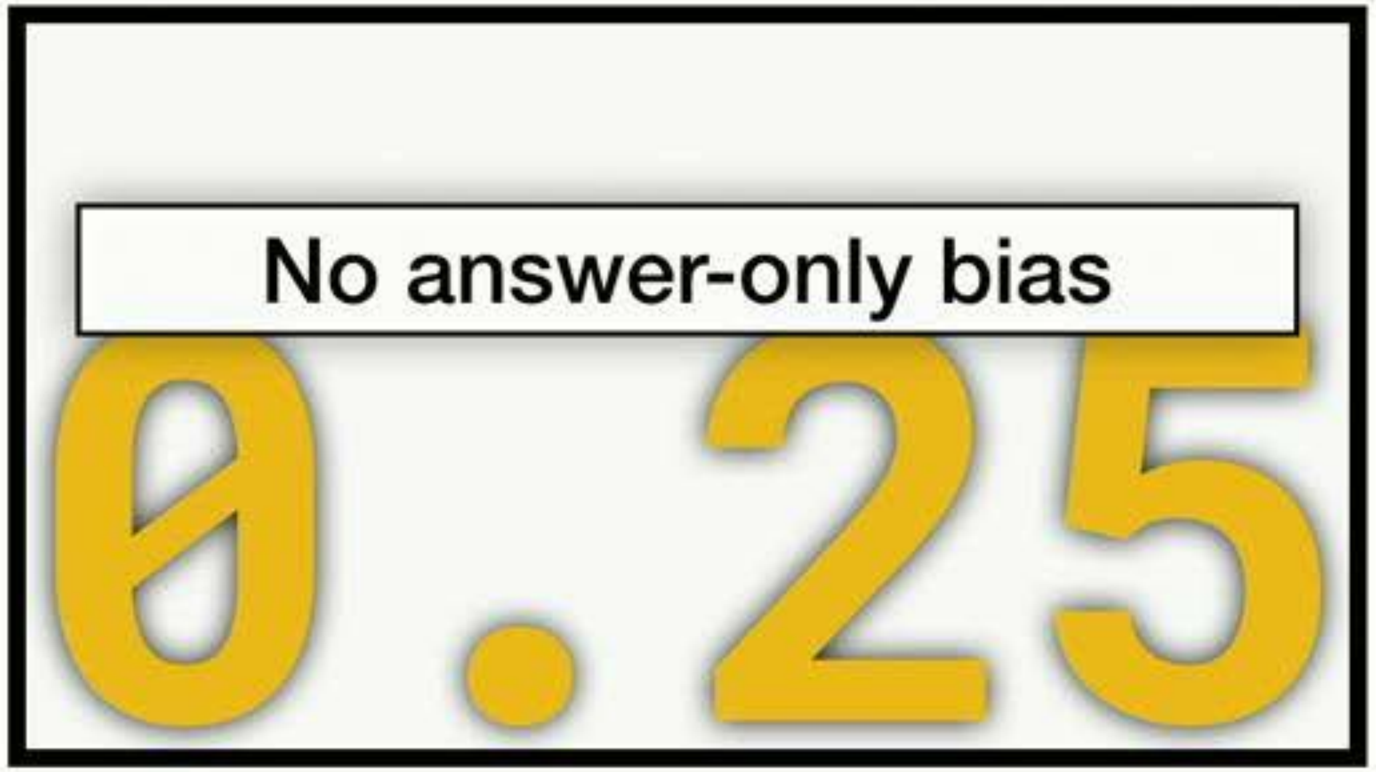
0.25

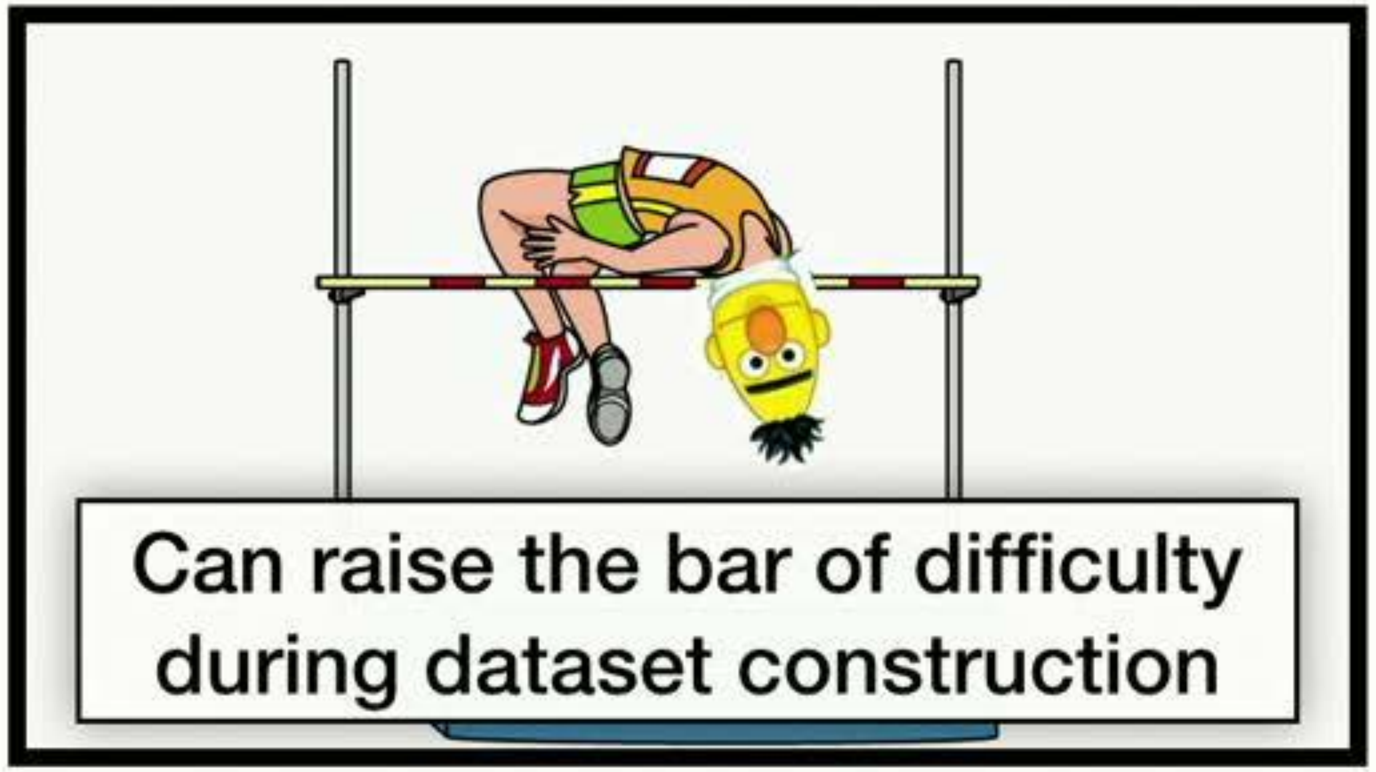
Unique Contributions of Adversarial Matching





Unique Contributions of Adversarial Matching





Putting it all together: VCR

- VCR features 290k questions over 110k images, each with answers and rationales.
- The questions are diverse and challenging.

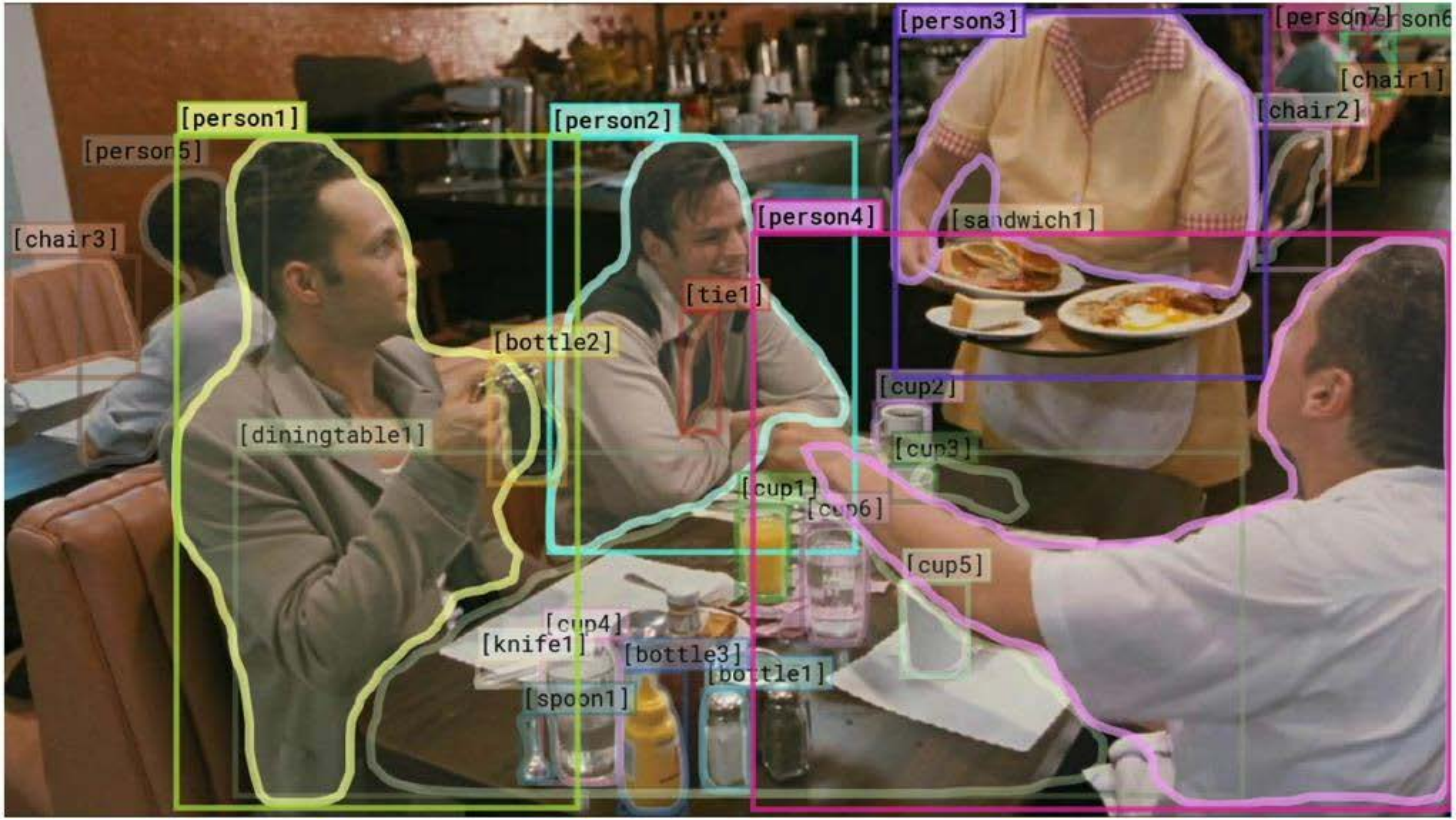
Our contributions

- New task: Visual Commonsense Reasoning

- Building VCR, feat. Adversarial Matching

- Recognition to Cognition Networks

Detour: Setting up the task



Detour: Setting up the task

Question answering

Answer justification

Detour: Setting up the task

Question answering

$$Q \rightarrow A$$

Answer justification

$$QA \rightarrow R$$

Detour: Setting up the task

Question answering

$$\underbrace{Q}_{\text{Query}} \rightarrow A$$

Answer justification

$$\underbrace{QA}_{\text{Query}} \rightarrow R$$

Let's make both tasks have the same format!

Detour: Setting up the task

Question answering

$$Q \rightarrow A$$

Query **Responses**

Answer justification

$$QA \rightarrow R$$

Query **Responses**

Let's make both tasks have the same format!

Detour: Setting up the task

Question answering

$$\underbrace{Q}_{\text{Query}} \rightarrow \underbrace{A}_{\text{Responses}}$$

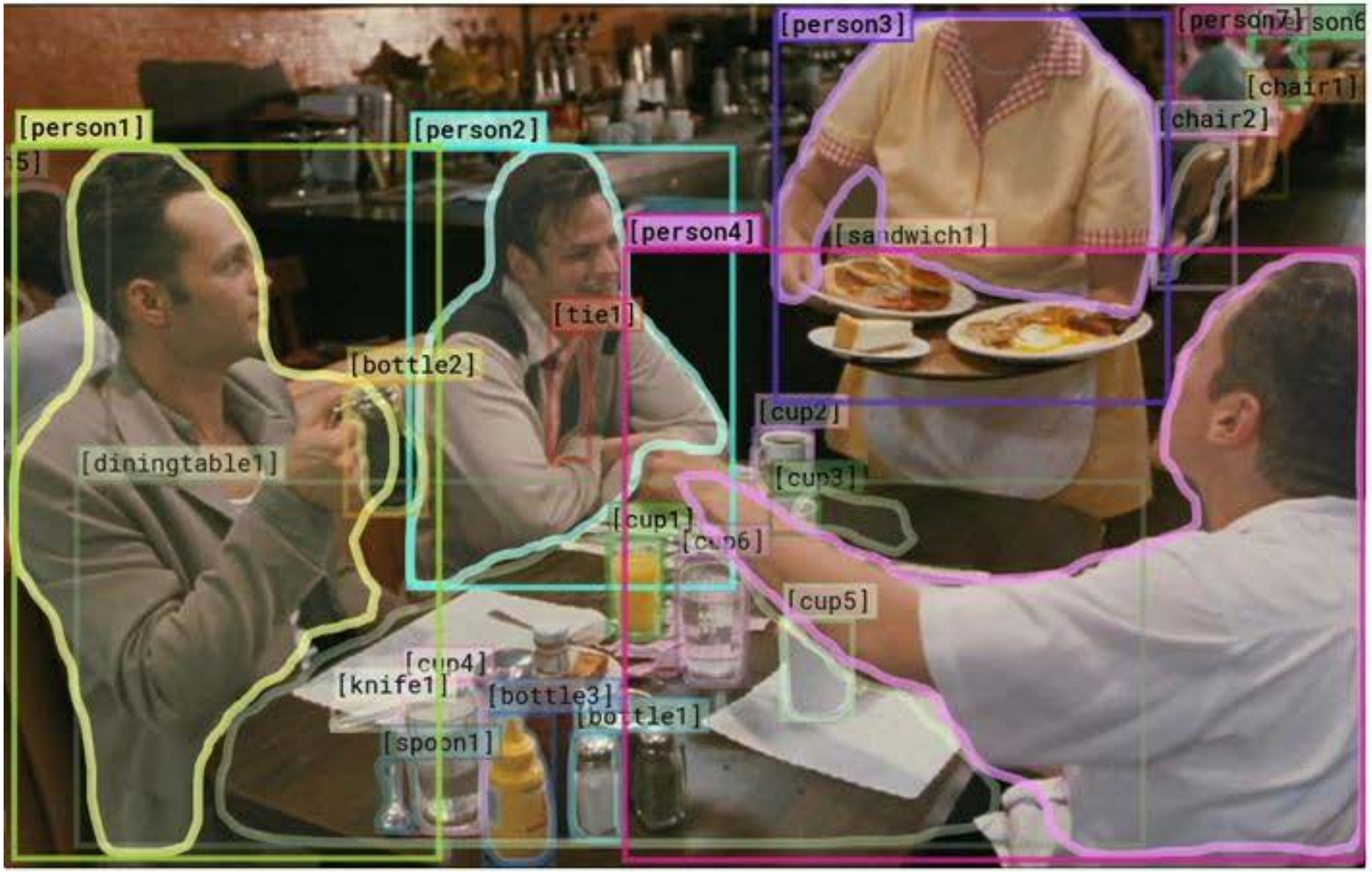
Answer justification

$$\underbrace{QA}_{\text{Query}} \rightarrow \underbrace{R}_{\text{Responses}}$$

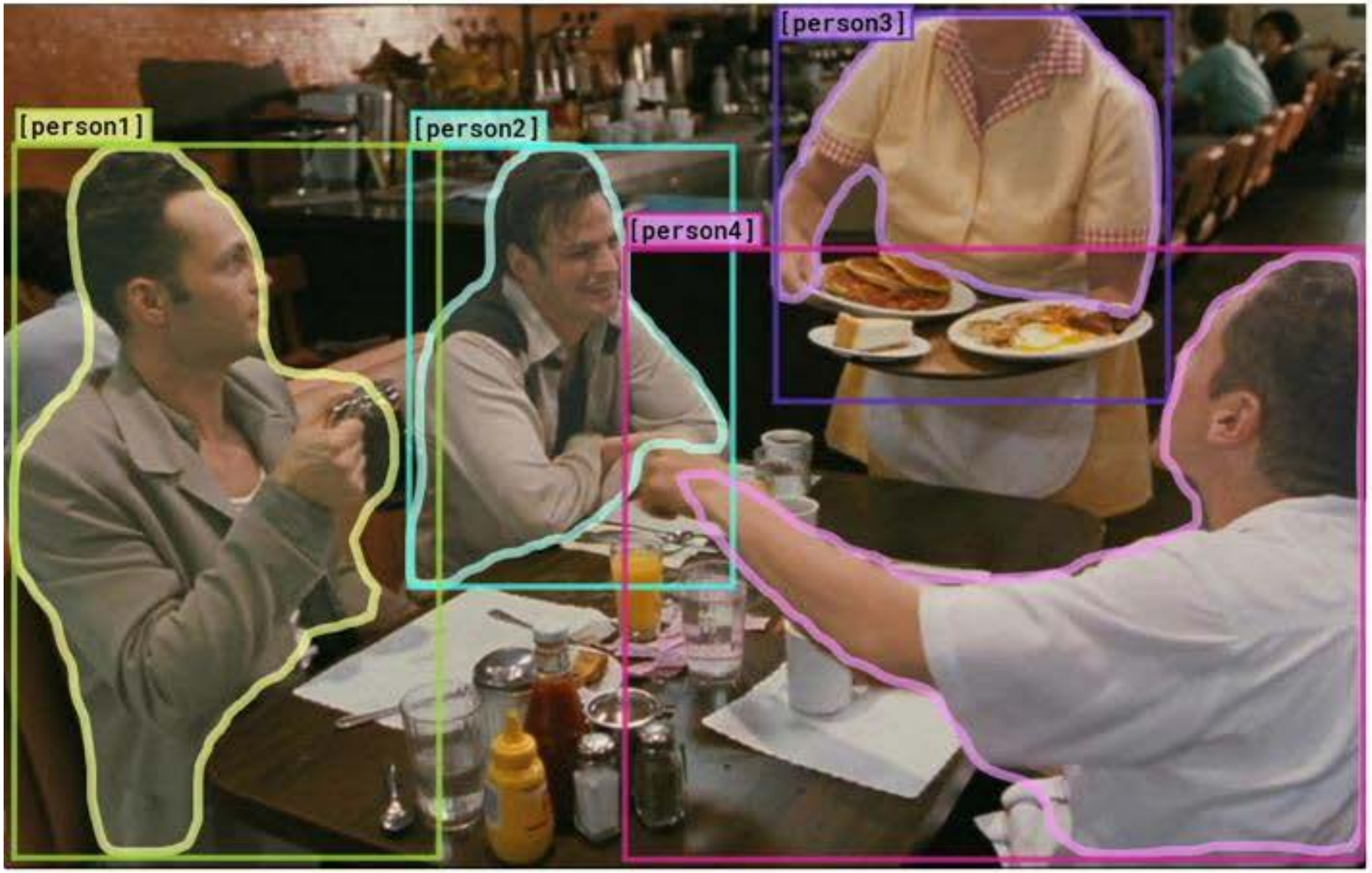
$$Q \rightarrow AR$$

Recognition to Cognition Networks

Recognition to Cognition Networks



Recognition to Cognition Networks



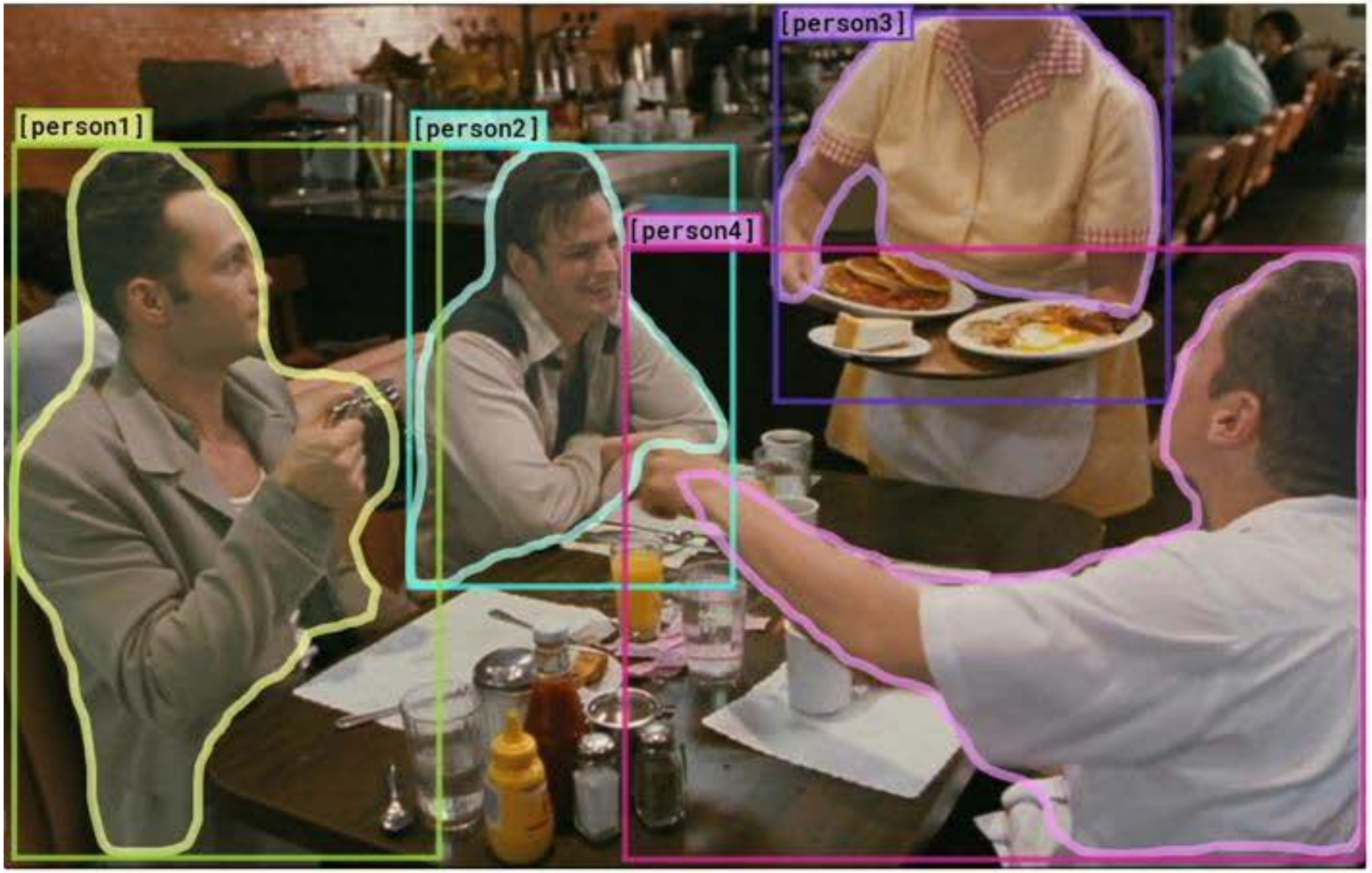
Query

Why is [person4 ]

pointing at [person1 ]?

Objects  ...

Recognition to Cognition Networks



Objects

...

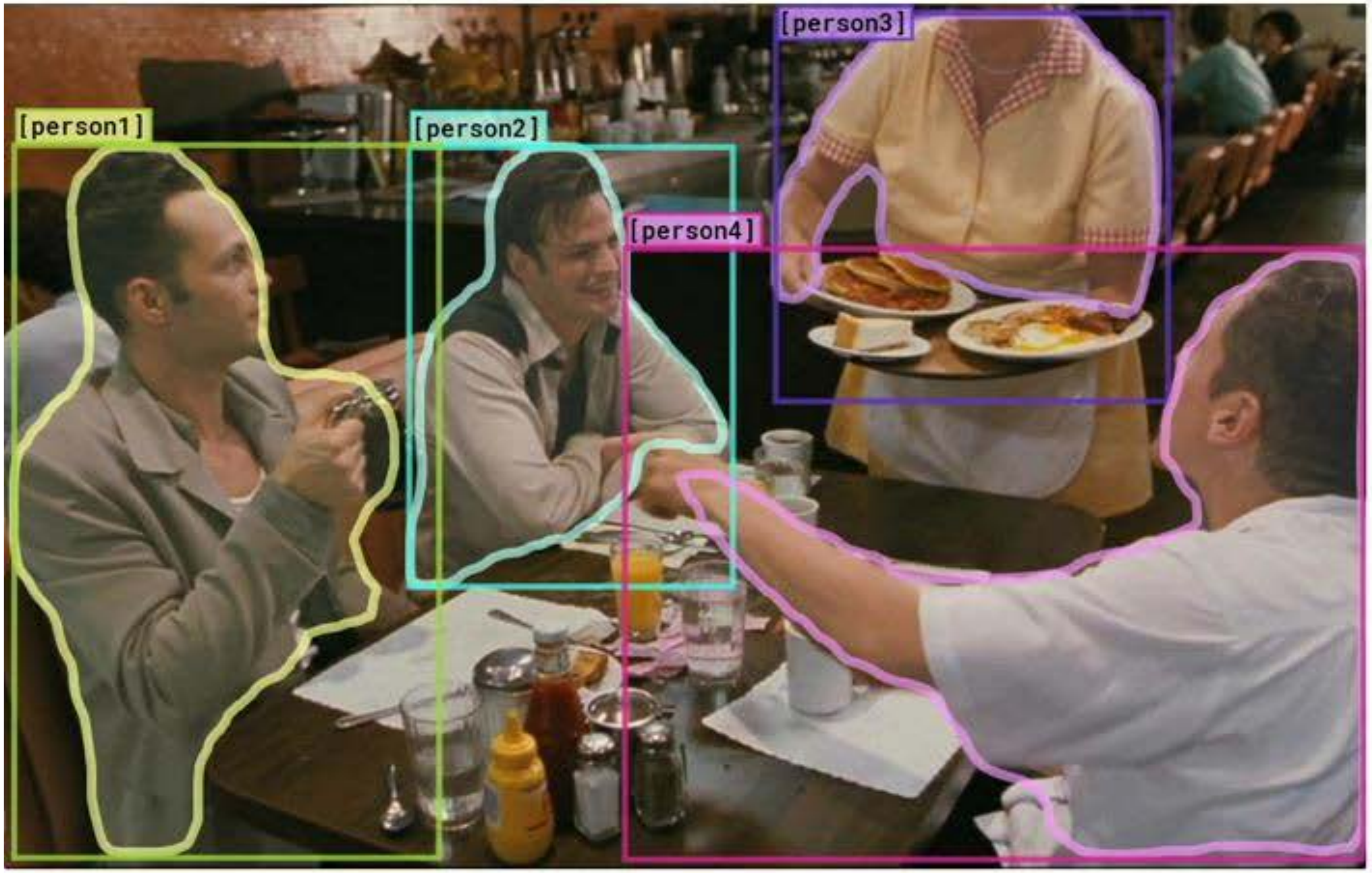
Query

Why is [person4] pointing at [person1]?

Response Choice

He is telling [person3] that [person1] ordered pancakes.

Recognition to Cognition Networks



Objects  ...

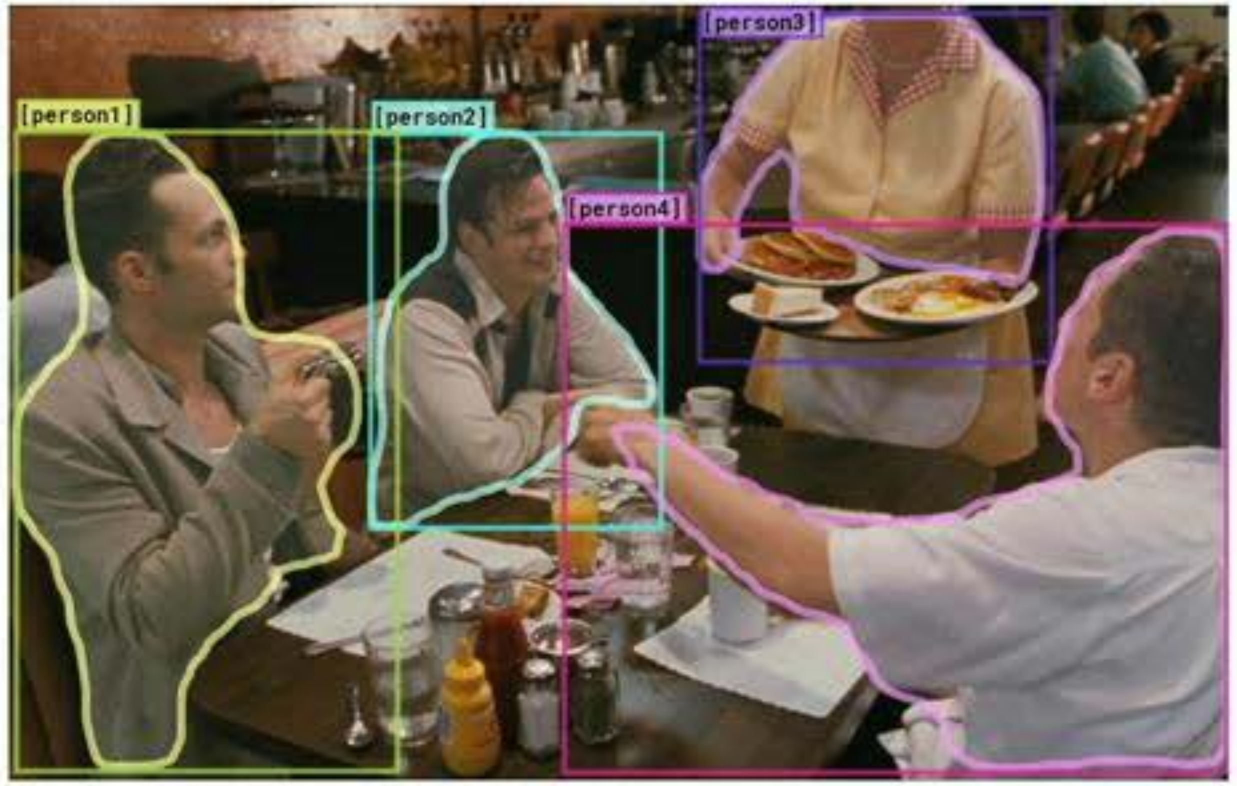
Query

Why is [person4] pointing at [person1]?

Response Choice

He is telling [person3] that [person1] ordered pancakes.

1. Grounding
2. Contextualization
3. Reasoning



Part 1: Grounding

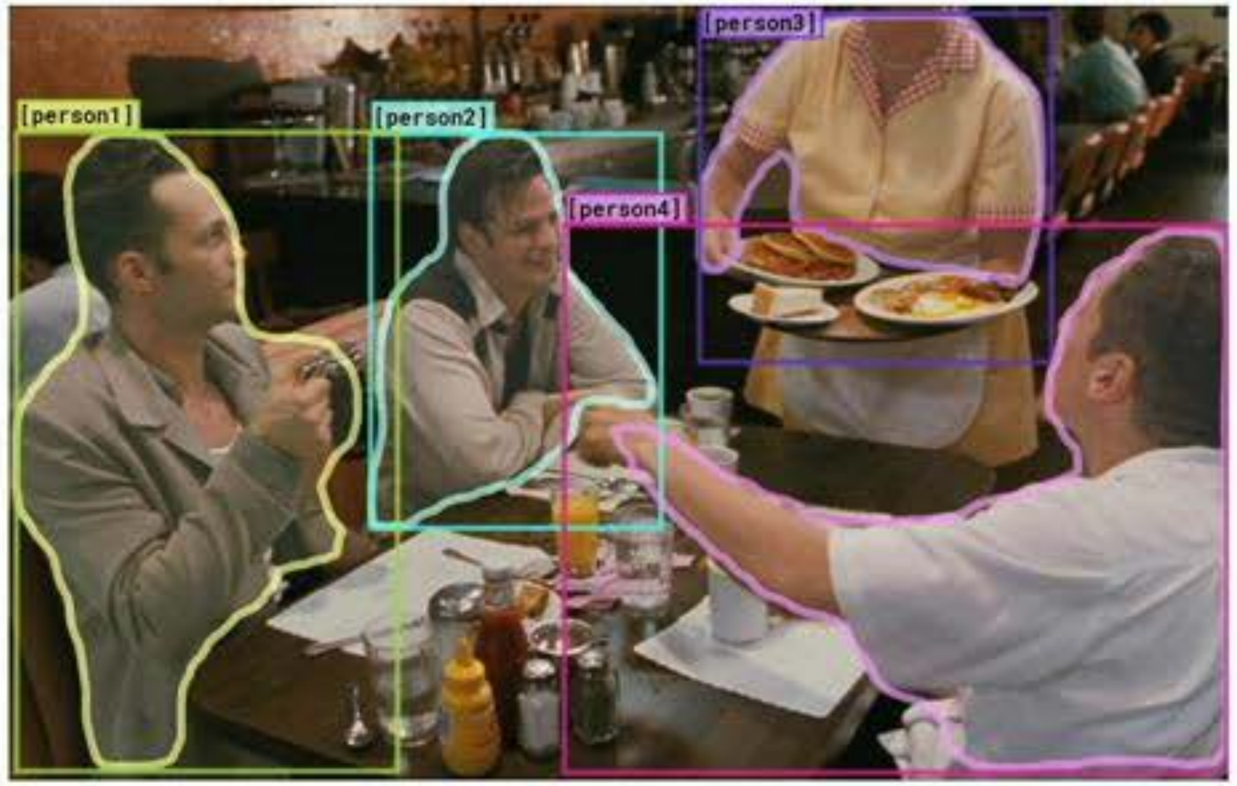
Objects  . . .

Query

Why is [person4] pointing at [person1]?

Response Choice

He is telling [person3] that [person1] ordered pancakes.



Part 1: Grounding

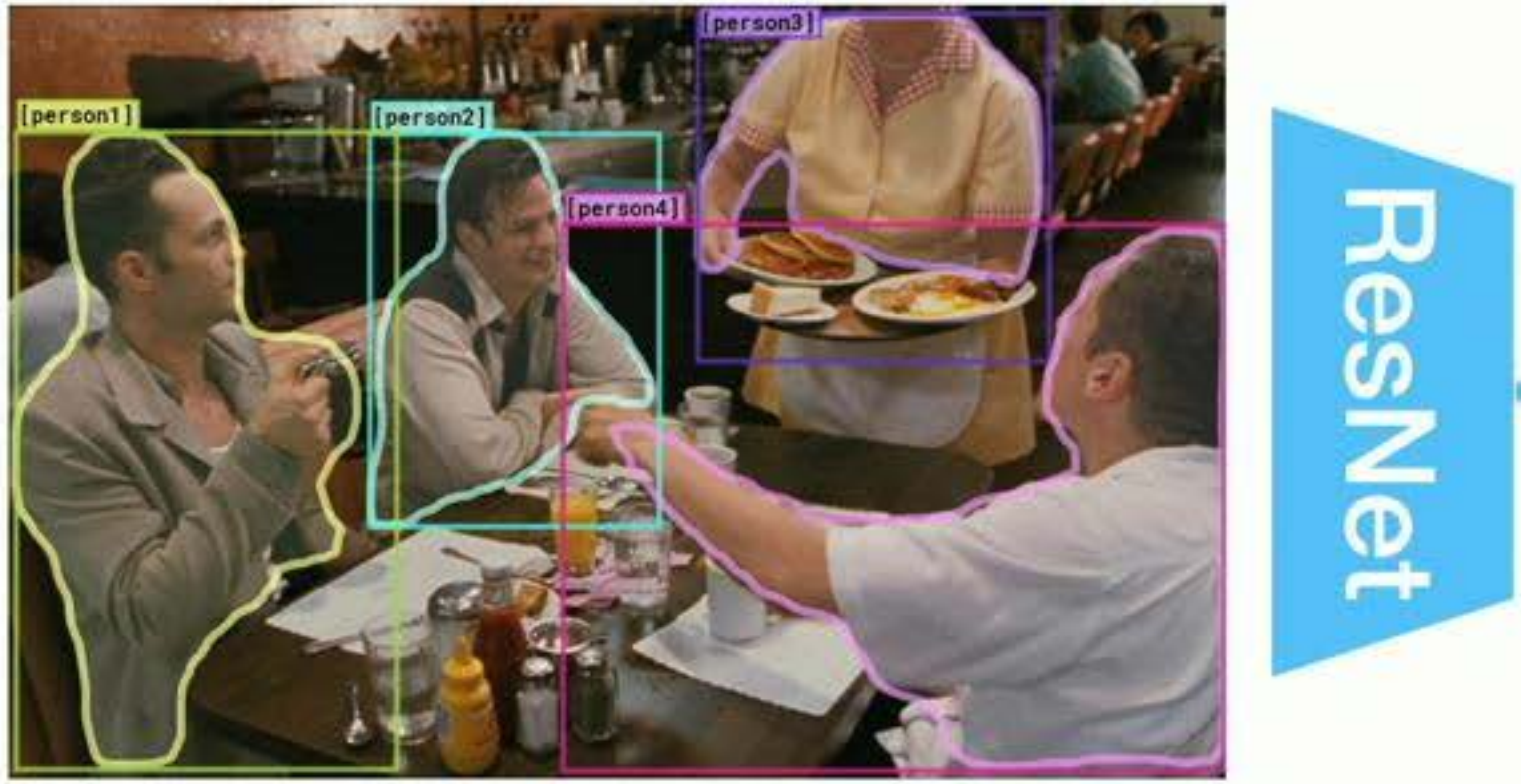
Objects p1 p2 p2 p3 ...

Query

Why is [person4]
pointing at [person1]?

Response Choice

He is telling [person3]
that [person1]
ordered pancakes.



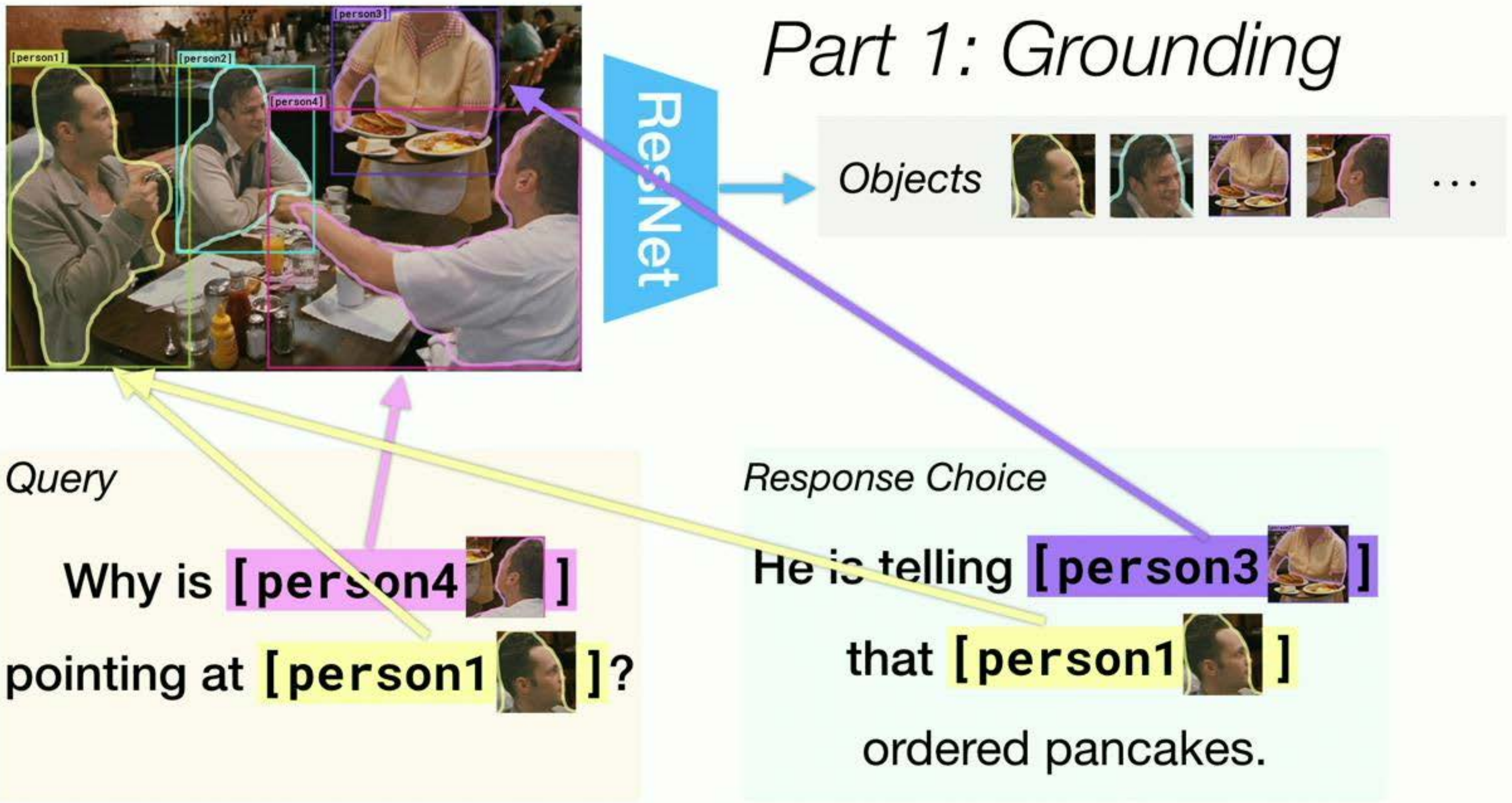
Part 1: Grounding

Query

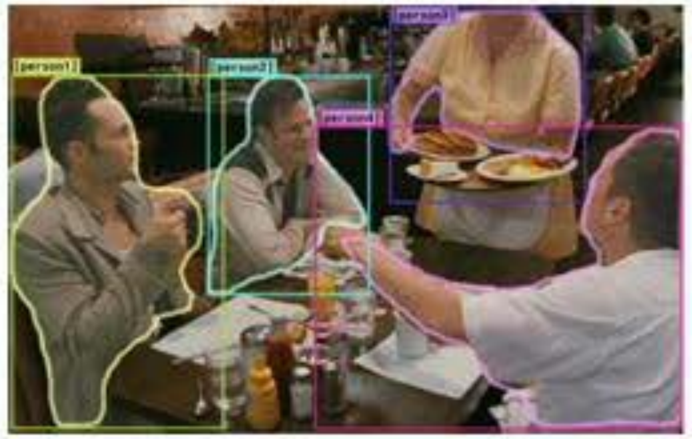
Why is [person4]
pointing at [person1]?

Response Choice

He is telling [person3]
that [person1]
ordered pancakes.



Part 1: Grounding
[person1]
[person2]
[person3]
[person4]
ResNet
Objects
...
Query
Why is [person4]
pointing at [person1]?
Response Choice
He is telling [person3]
that [person1]
ordered pancakes.



Objects  ...

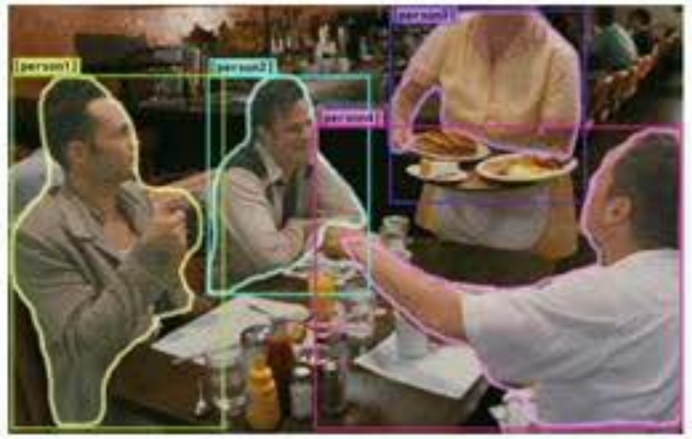
Part 1: Grounding

Query

Why is [person4] pointing at [person1]?

Response Choice

He is telling [person3] that [person1] ordered pancakes.



Objects  ...

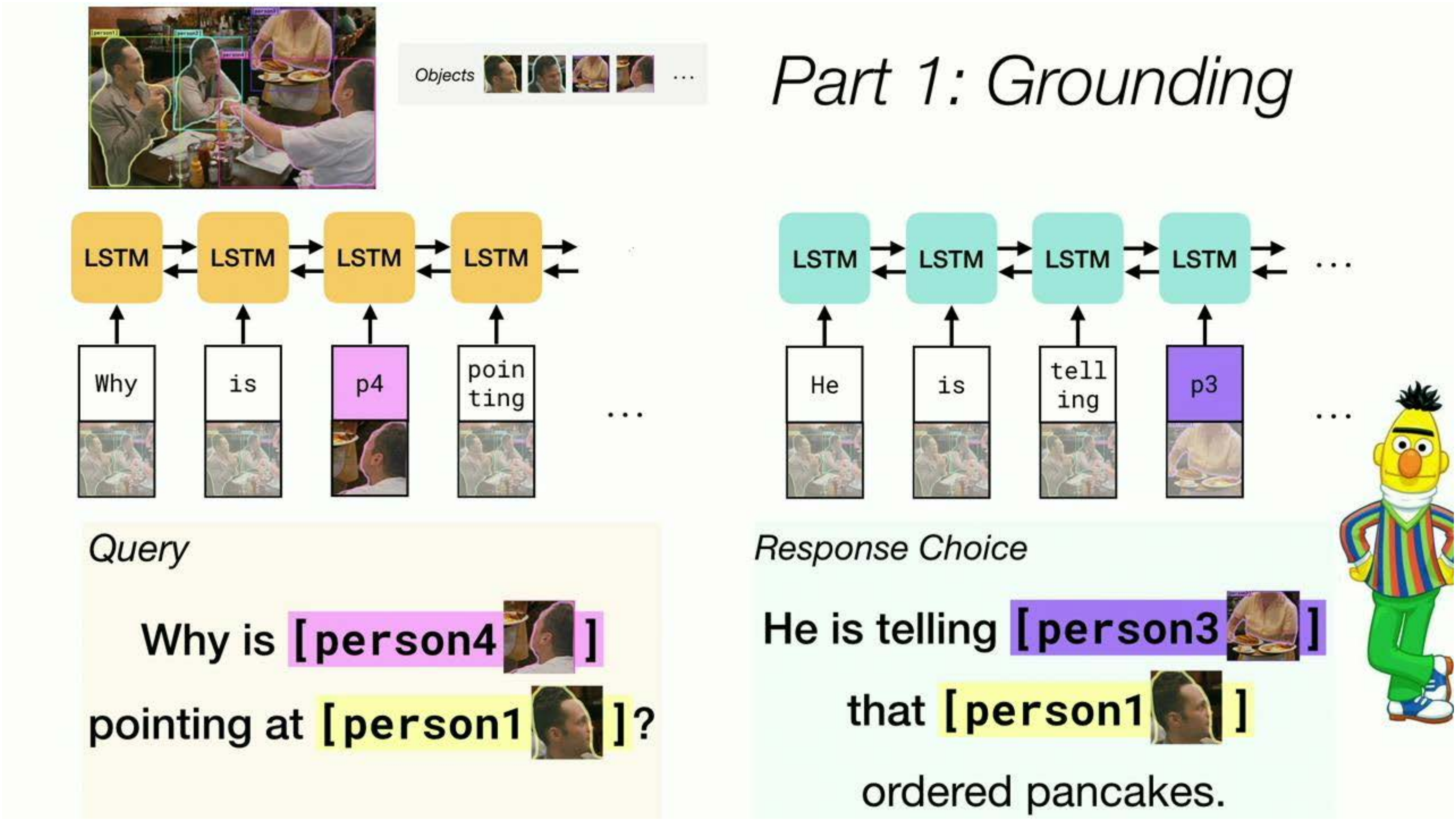
Part 1: Grounding

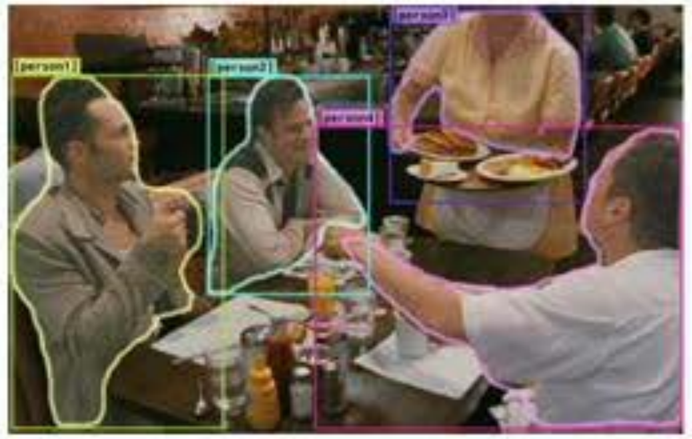
Query

Why is [person4  **]**
pointing at [person1  **]?**

Response Choice

He is telling [person3  **]**
that [person1  **]**
ordered pancakes.


Objects ...
Part 1: Grounding
LSTM
LSTM
LSTM
LSTM
Why
is
p4
poin
ting
...
LSTM
LSTM
LSTM
LSTM
...
He
is
tell
ing
p3
...
Query
Why is [person4]
pointing at [person1]?
Response Choice
He is telling [person3]
that [person1]
ordered pancakes.



Part 2: Contextualization

Objects  ...

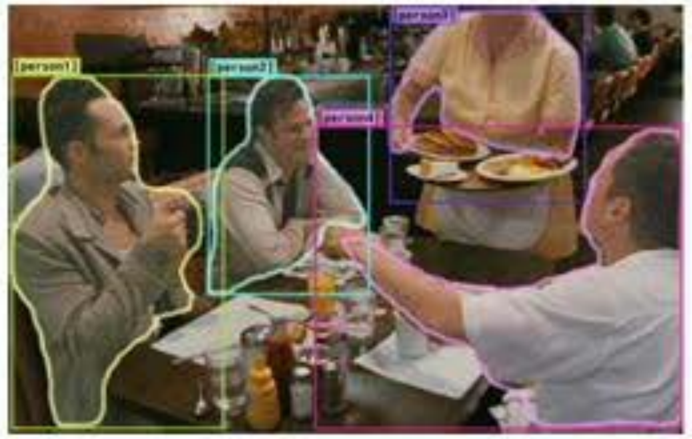
Query

Why is [person4 ] pointing at [person1 ]?

Response Choice

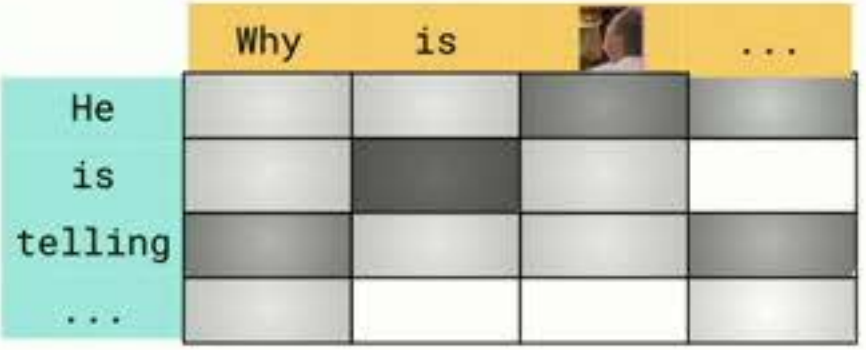
He is telling [person3 ] that [person1 ] ordered pancakes.

Part 2: Contextualization



Objects

...

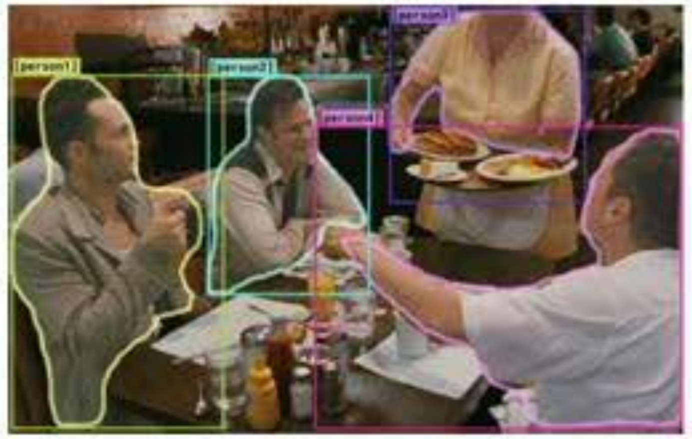


Query

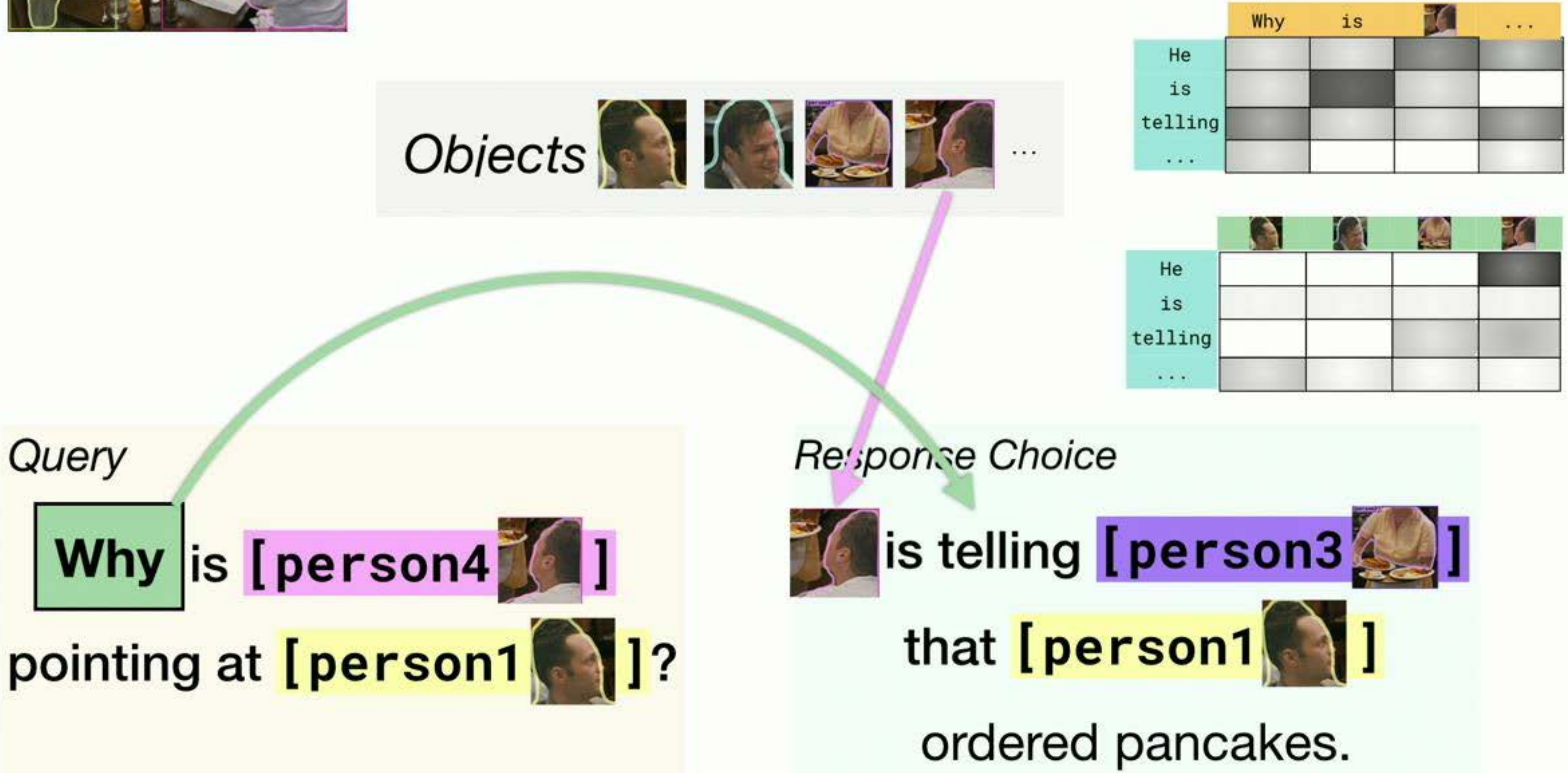
Why is [person4] pointing at [person1]?

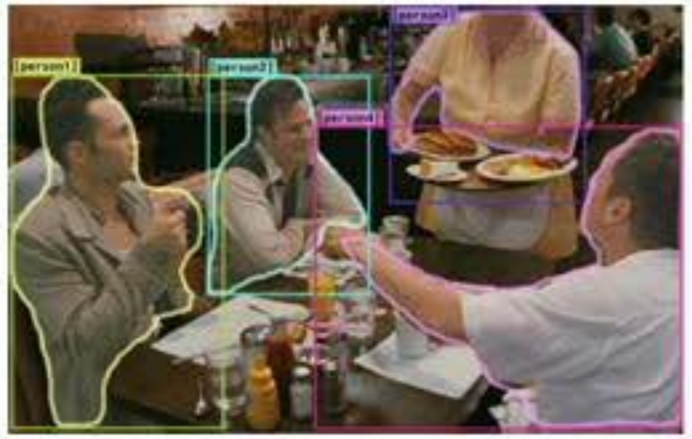
Response Choice

He is telling [person3] that [person1] ordered pancakes.



Part 2: Contextualization

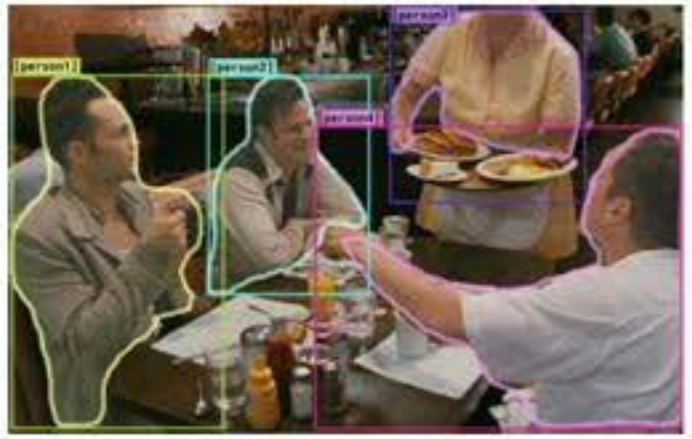




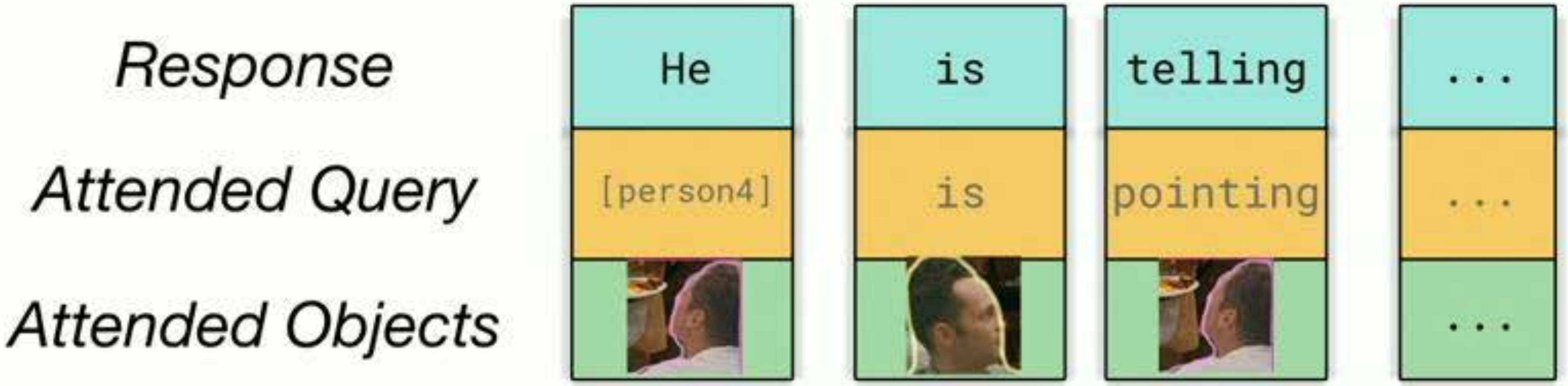
Part 3: Reasoning

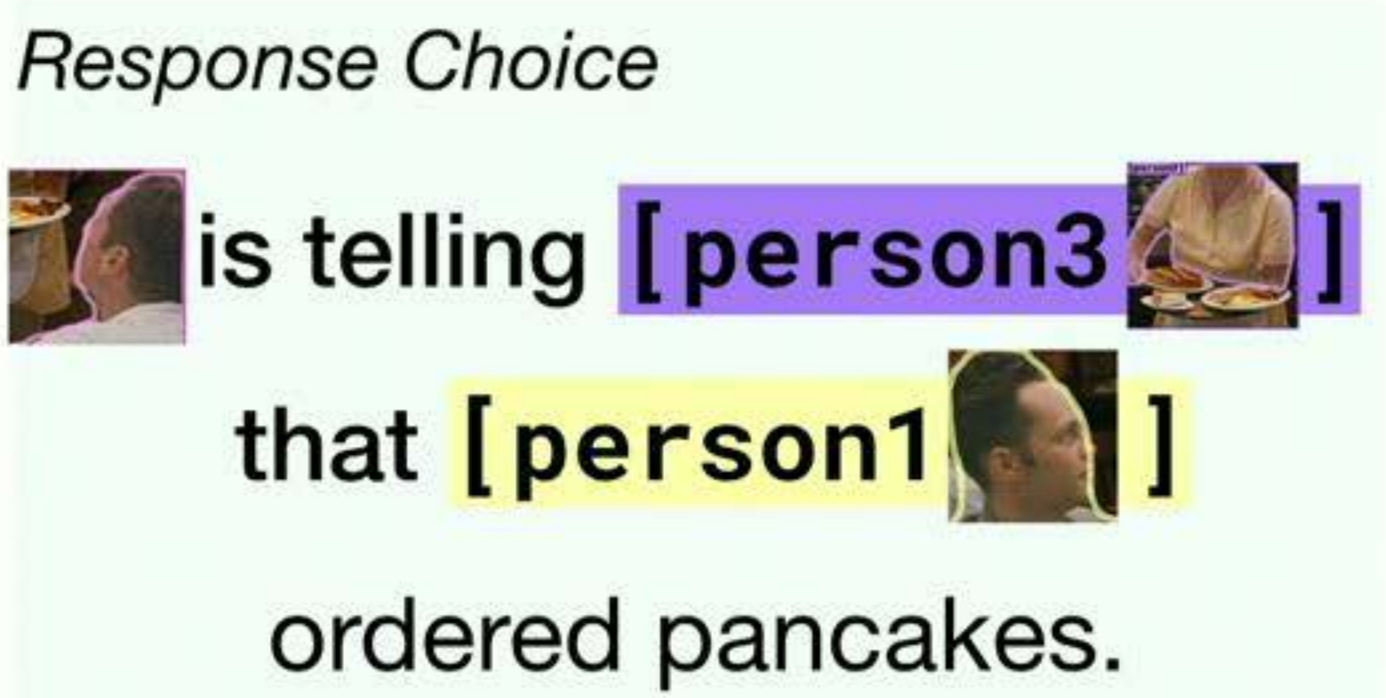
Response Choice

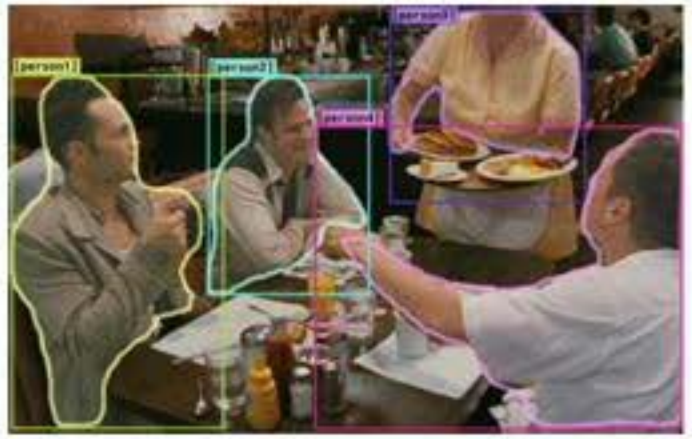
 is telling **[person3**  **]** that **[person1**  **]** ordered pancakes.



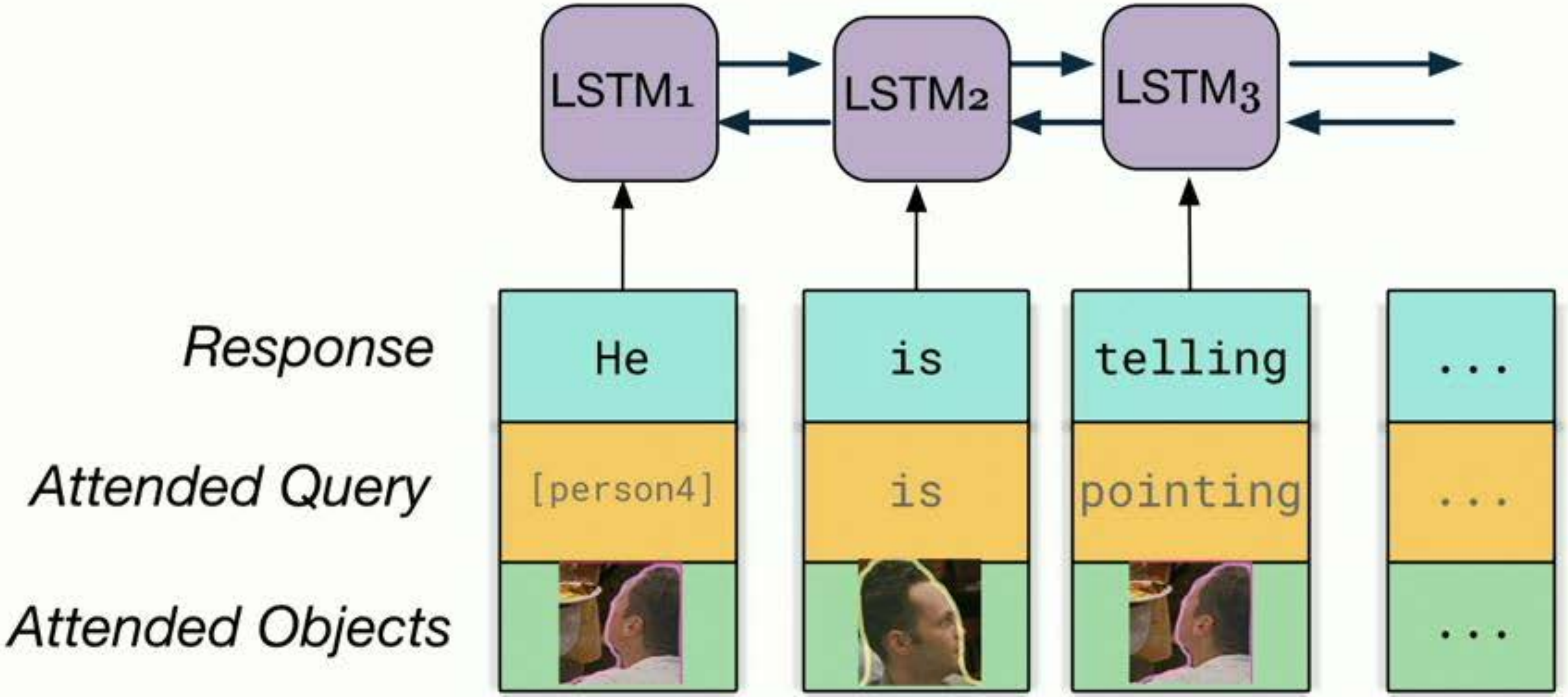
Part 3: Reasoning

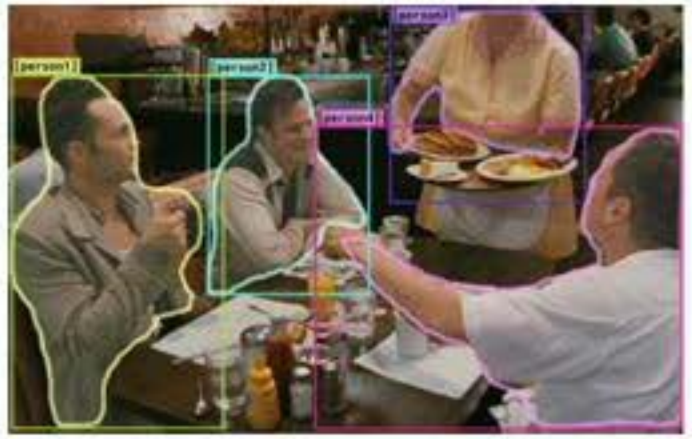




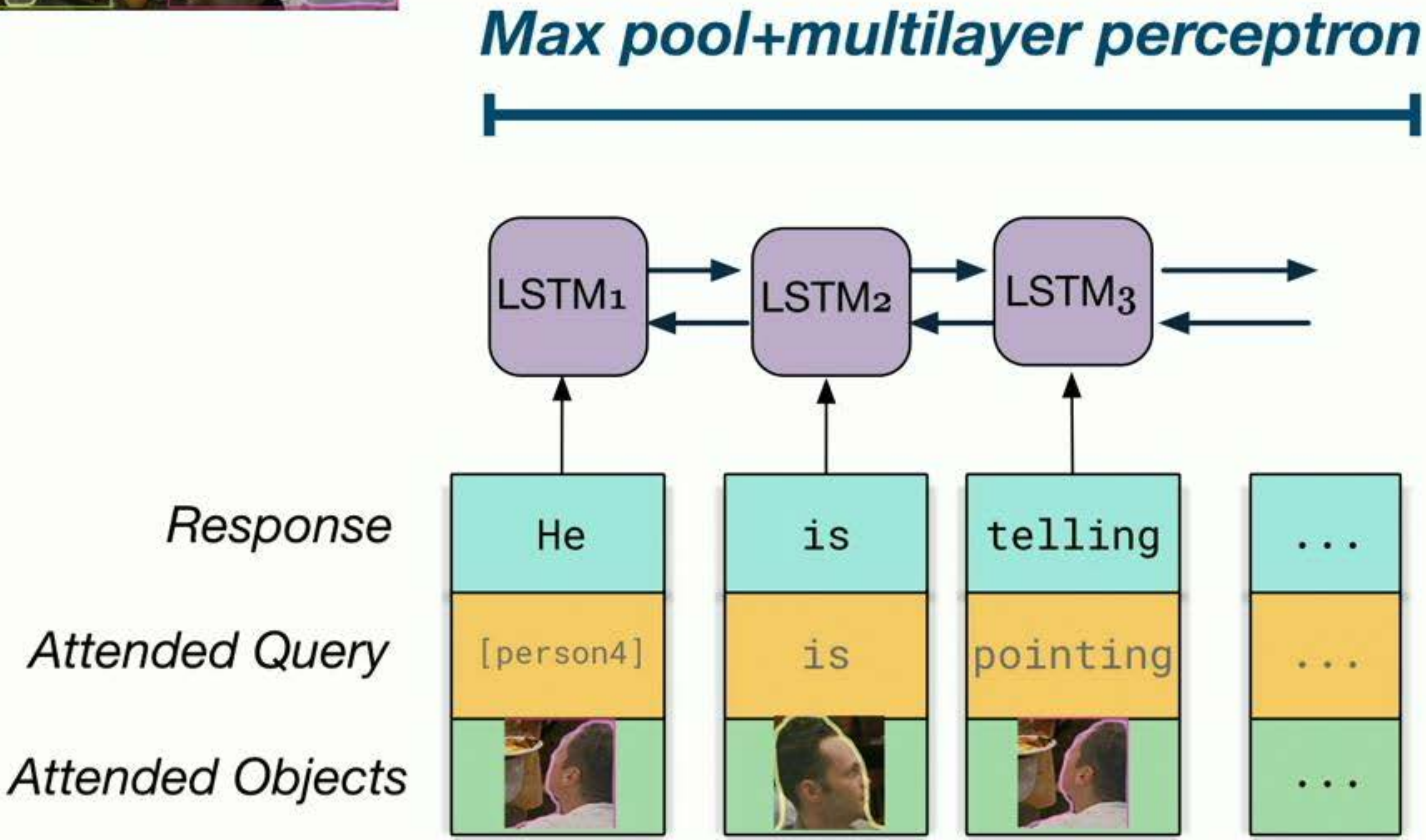


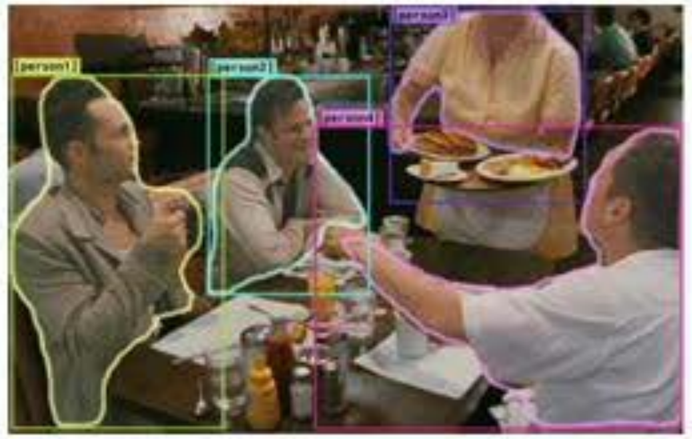
Part 3: Reasoning



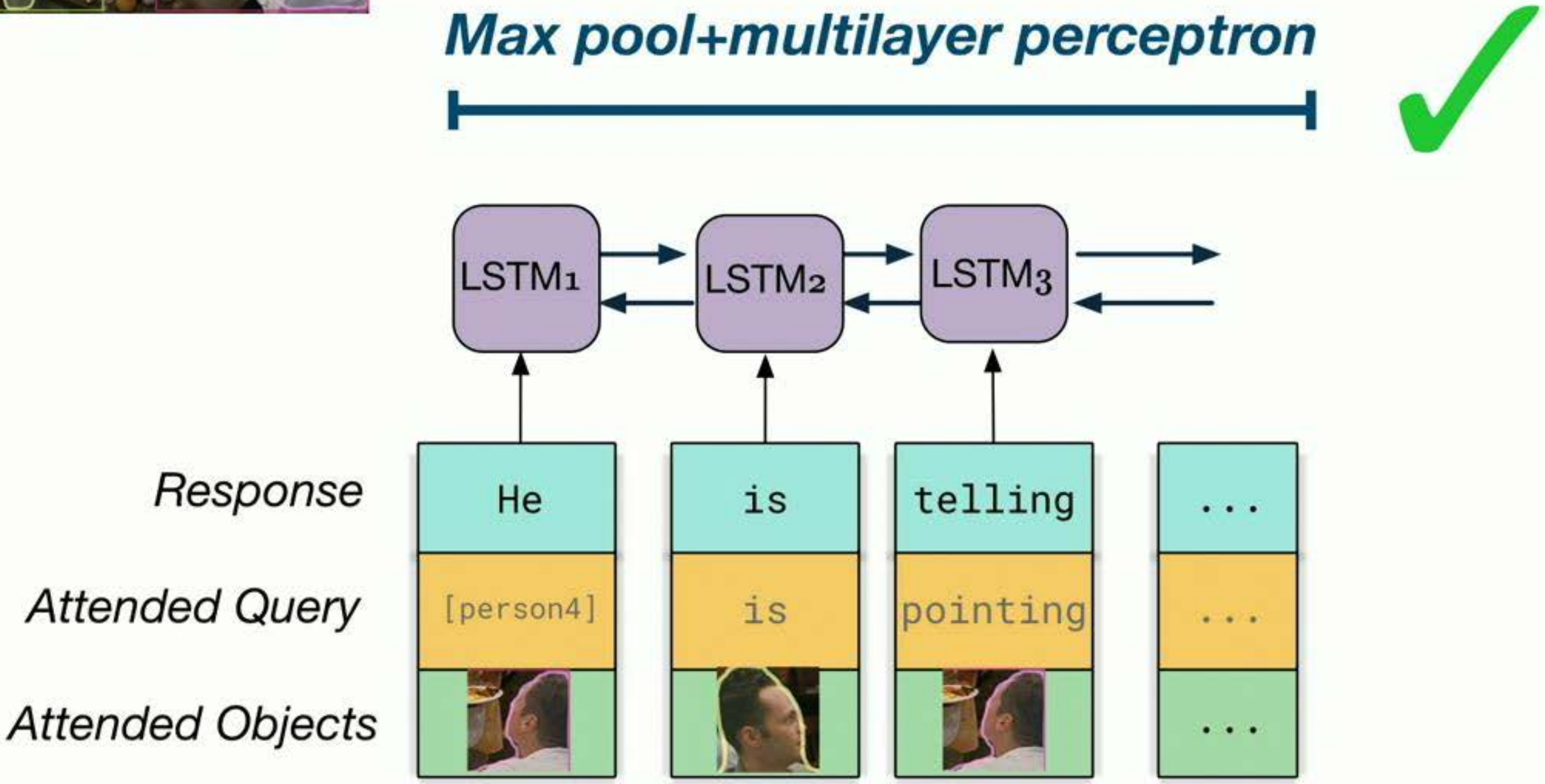


Part 3: Reasoning





Part 3: Reasoning



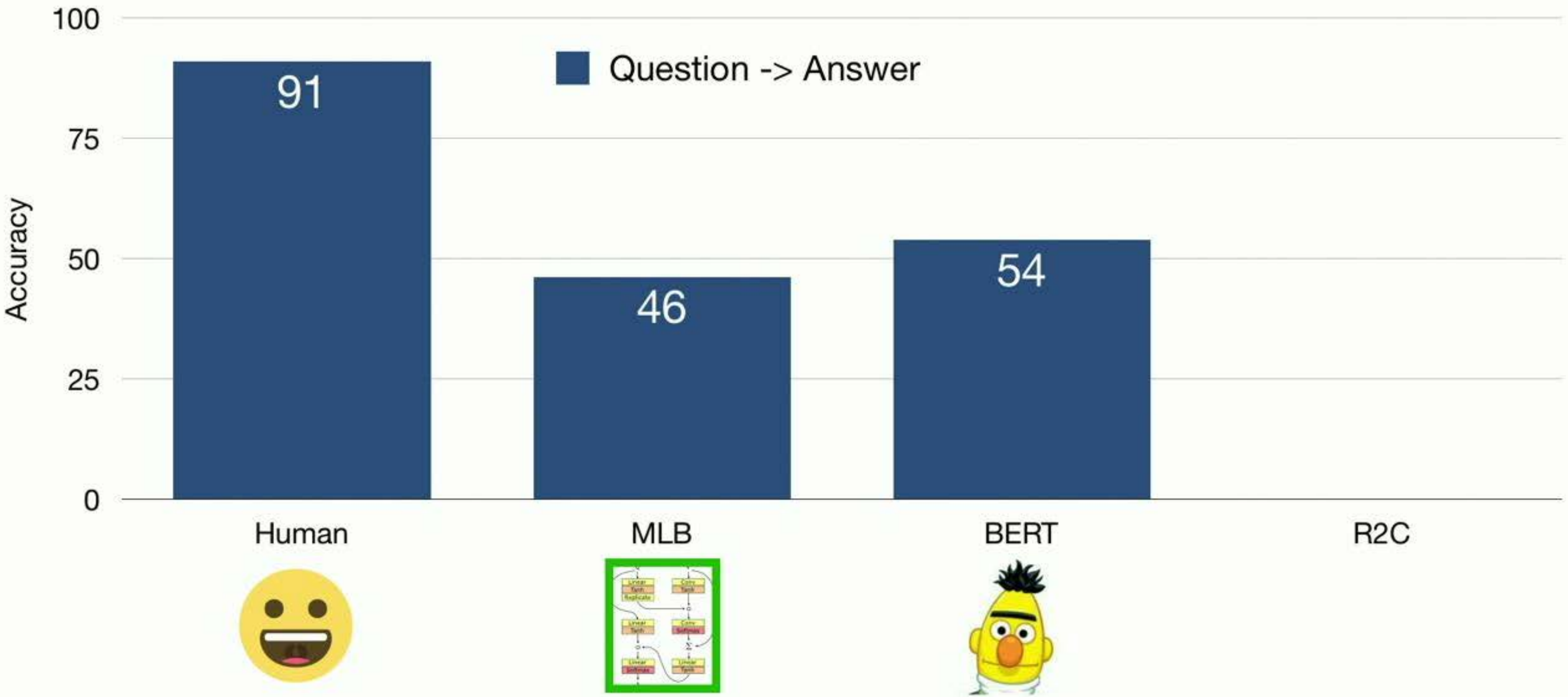
VCR Results

VCR Results

VCR Results

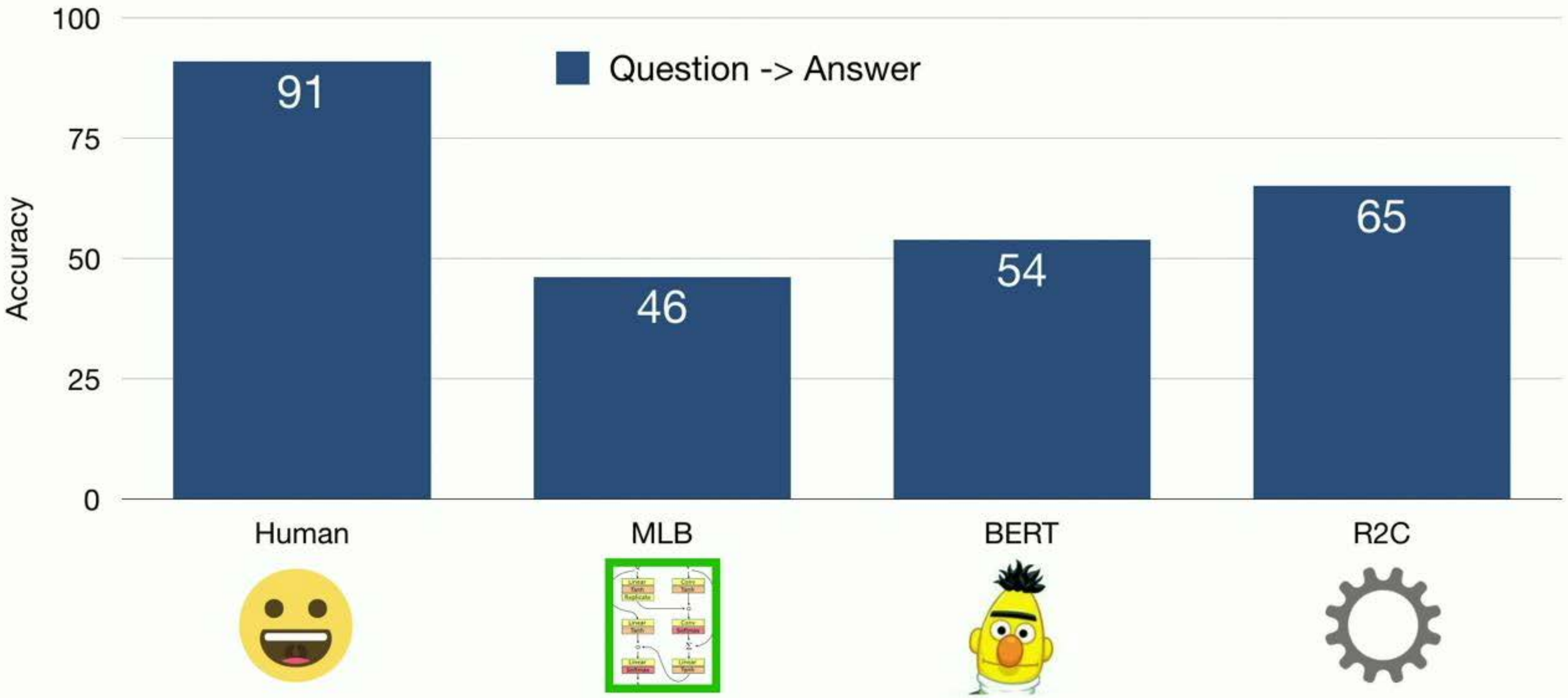

Accuracy
100
75
50
25
0
Question -> Answer
91
46
Human
MLB
BERT
R2C

VCR Results



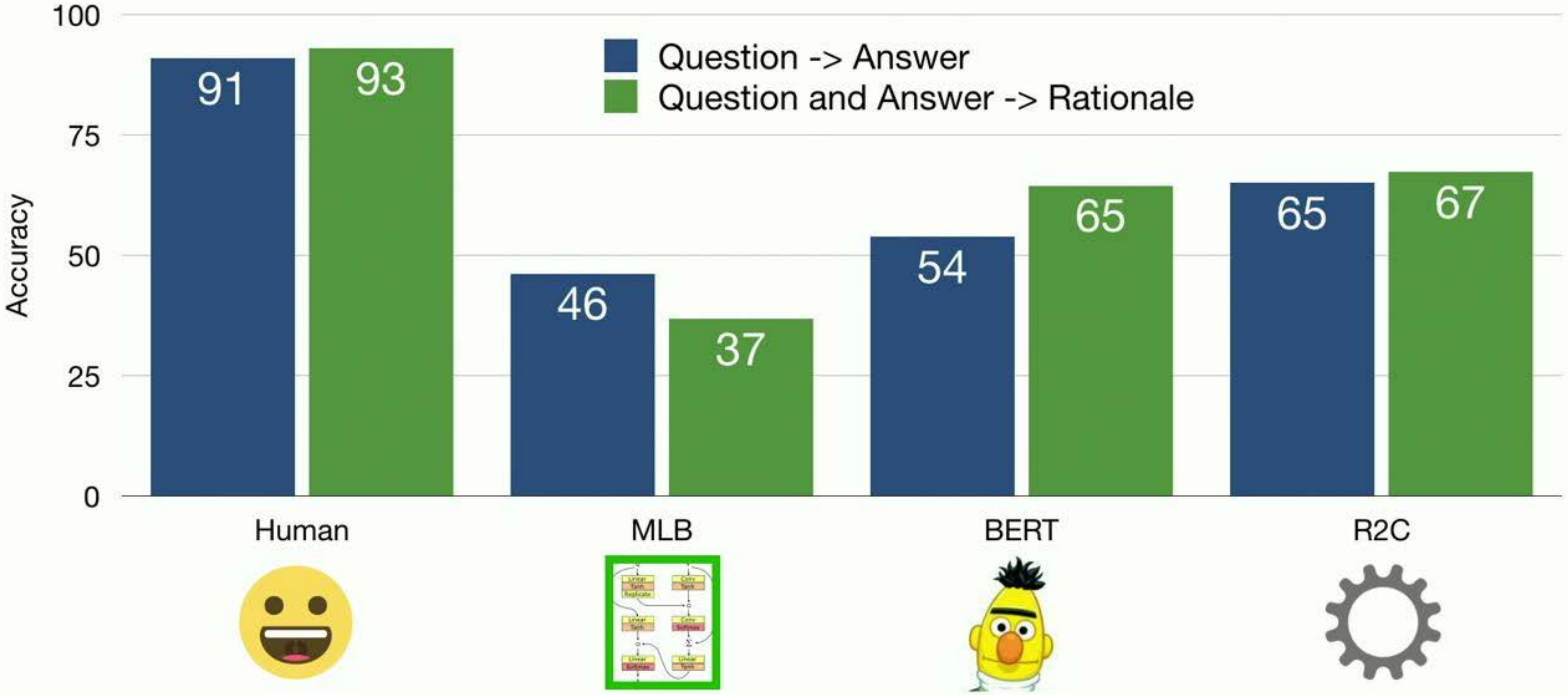
100
75
50
25
0
Accuracy
Question -> Answer
91
46
54
Human
MLB
BERT
R2C

VCR Results

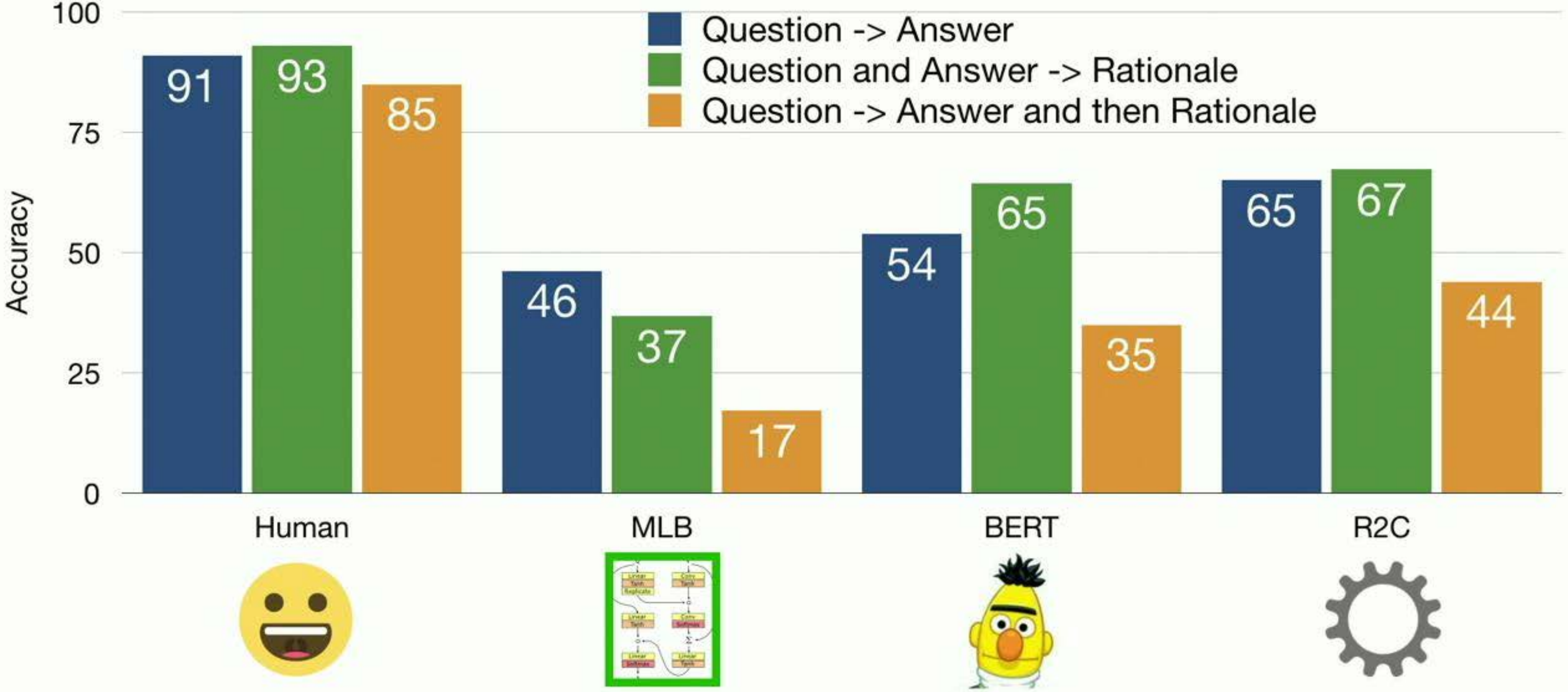


Question -> Answer
Accuracy
100
75
50
25
0
91
46
54
65
Human
MLB
BERT
R2C

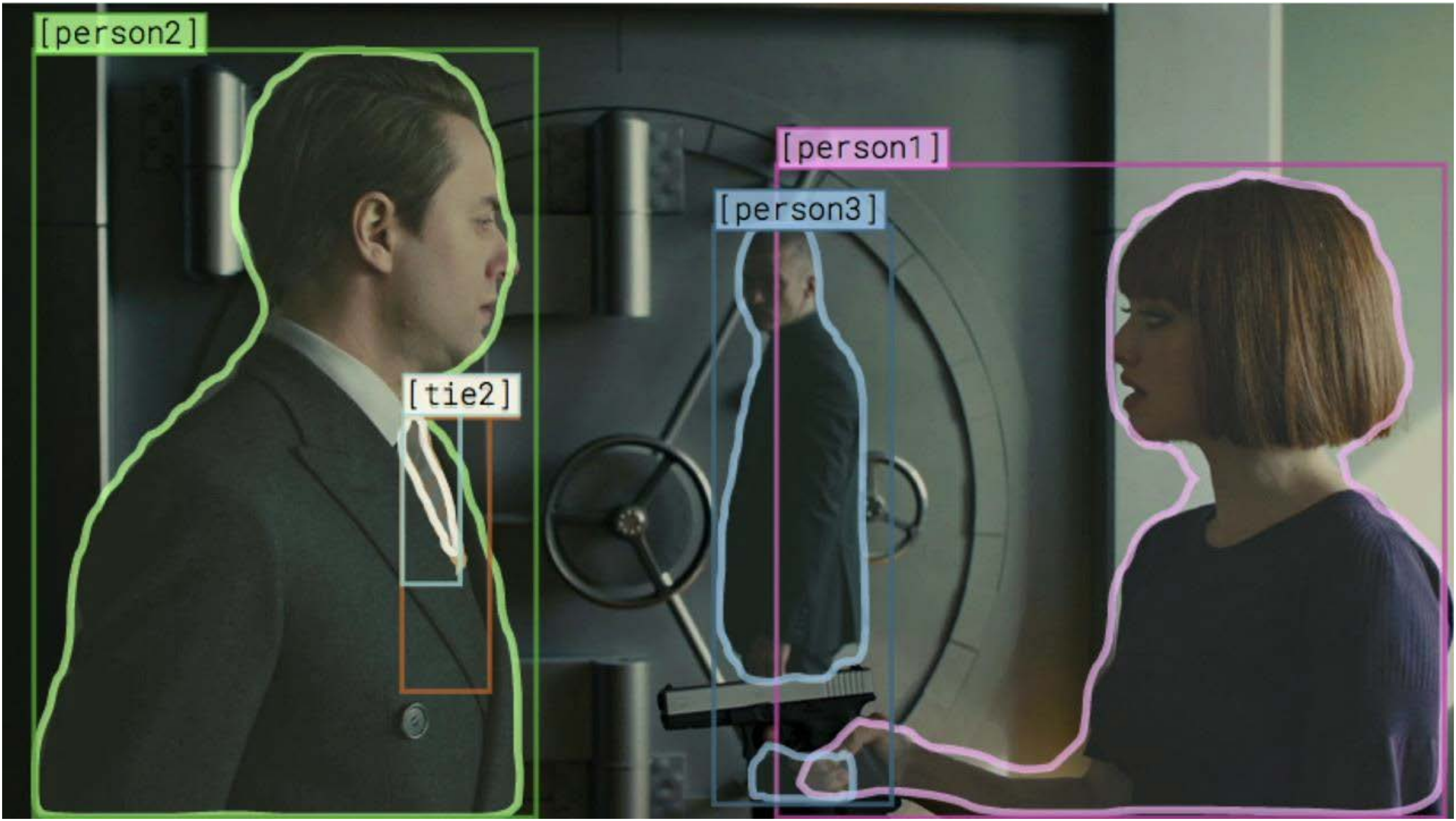
VCR Results

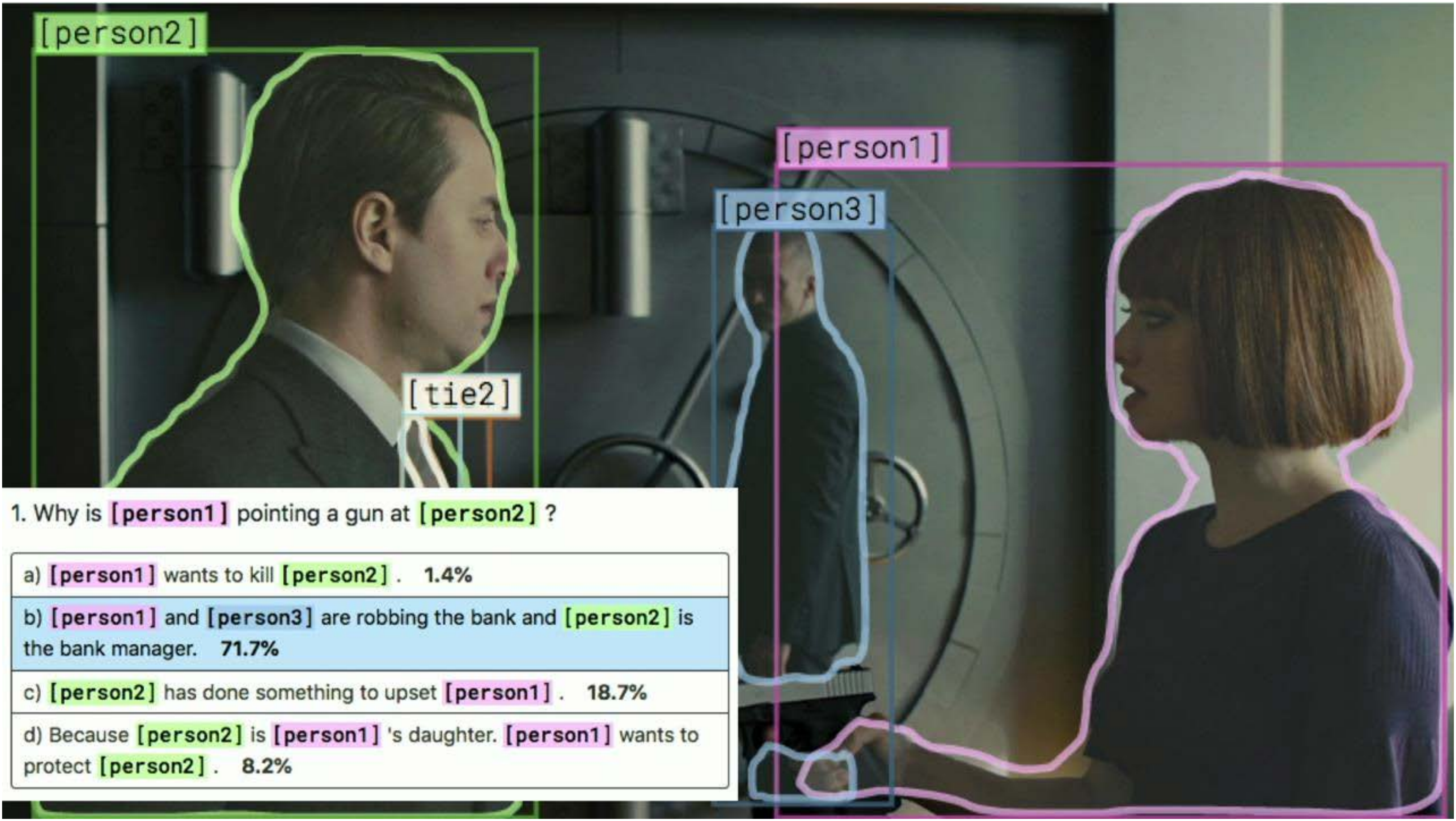

Accuracy
100
75
50
25
0
Question -> Answer
Question and Answer -> Rationale
91
93
46
37
54
65
65
67
Human
MLB
BERT
R2C

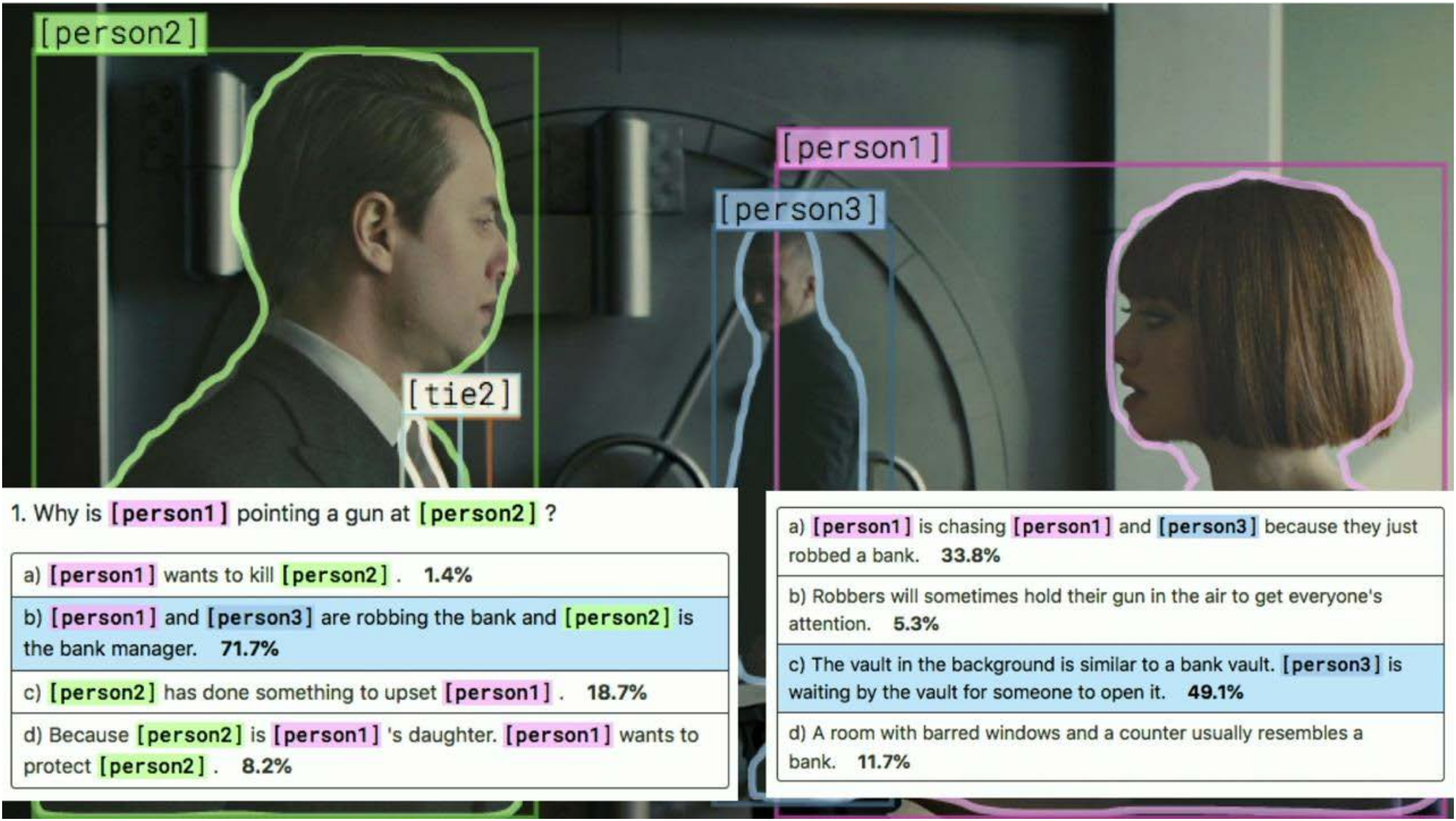
VCR Results



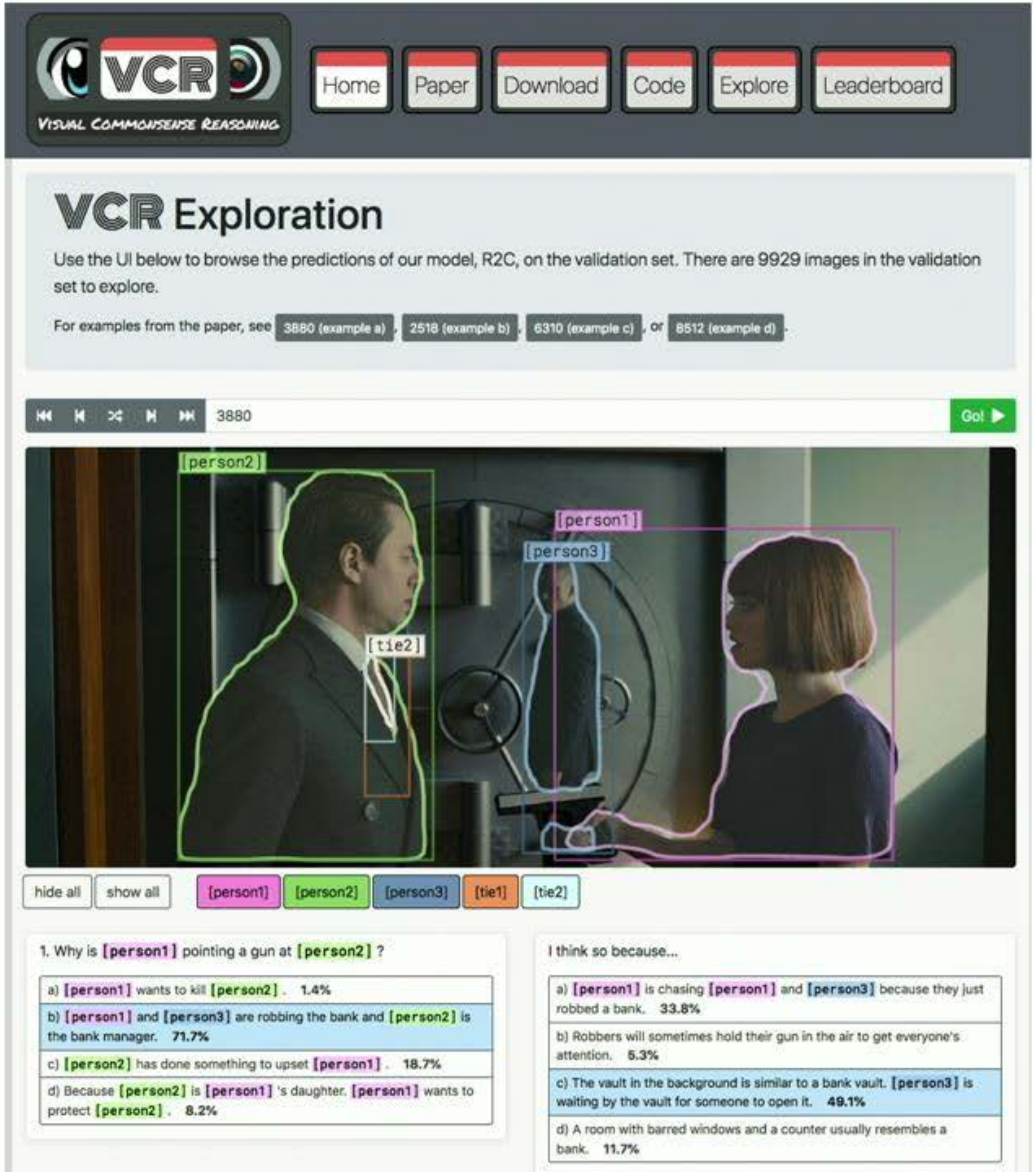
100
75
50
25
0
Accuracy
Question -> Answer
Question and Answer -> Rationale
Question -> Answer and then Rationale
91
93
85
46
37
17
54
65
35
65
67
44
Human
MLB
BERT
R2C


[person2]
[tie2]
[person3]
[person1]


[person2]
[person1]
[person3]
[tie2]
1. Why is [person1] pointing a gun at [person2] ?
a) [person1] wants to kill [person2] . 1.4%
b) [person1] and [person3] are robbing the bank and [person2] is the bank manager. 71.7%
c) [person2] has done something to upset [person1] . 18.7%
d) Because [person2] is [person1] 's daughter. [person1] wants to protect [person2] . 8.2%


[person2]
[person1]
[person3]
[tie2]
1. Why is [person1] pointing a gun at [person2] ?
a) [person1] wants to kill [person2] . 1.4%
b) [person1] and [person3] are robbing the bank and [person2] is the bank manager. 71.7%
c) [person2] has done something to upset [person1] . 18.7%
d) Because [person2] is [person1] 's daughter. [person1] wants to protect [person2] . 8.2%
a) [person1] is chasing [person1] and [person3] because they just robbed a bank. 33.8%
b) Robbers will sometimes hold their gun in the air to get everyone's attention. 5.3%
c) The vault in the background is similar to a bank vault. [person3] is waiting by the vault for someone to open it. 49.1%
d) A room with barred windows and a counter usually resembles a bank. 11.7%

visualcommonsense.com



visualcommonsense.com

VCR
VISUAL COMMONSENSE REASONING

Home | Paper | Download | Code | Explore | Leaderboard

Visual Commonsense Reasoning

On VCR, a model must not only answer commonsense visual questions, but also provide a rationale that explains why the answer is true.

Read our paper on arXiv»

Submitting to the leaderboard

Submission is easy! You just need to email Rowan with your predictions. Formatting instructions are below:

Submit to the leaderboard»

What kinds of submissions are allowed?

The only constraint is that your system must predict the **answer first**, then the rationale. (The rationales were selected to be highly relevant to the correct Q,A pair, so they leak information about the correct answer.)

- To deter this, the submission format involves submitting predictions for each possible rationale, conditioned on each possible answer.
- A simple way of setting up the experiments (used in the paper) is to consider a task with *query* and four *response* choices. For Q->A the query is the question, and the response choices are the answers. For QA->R, the query is the question and answer, concatenated together, and the response choices are the rationales.

Questions?

If it's not about something private, check out the google group below:

Get help via the google group»

VCR Leaderboard

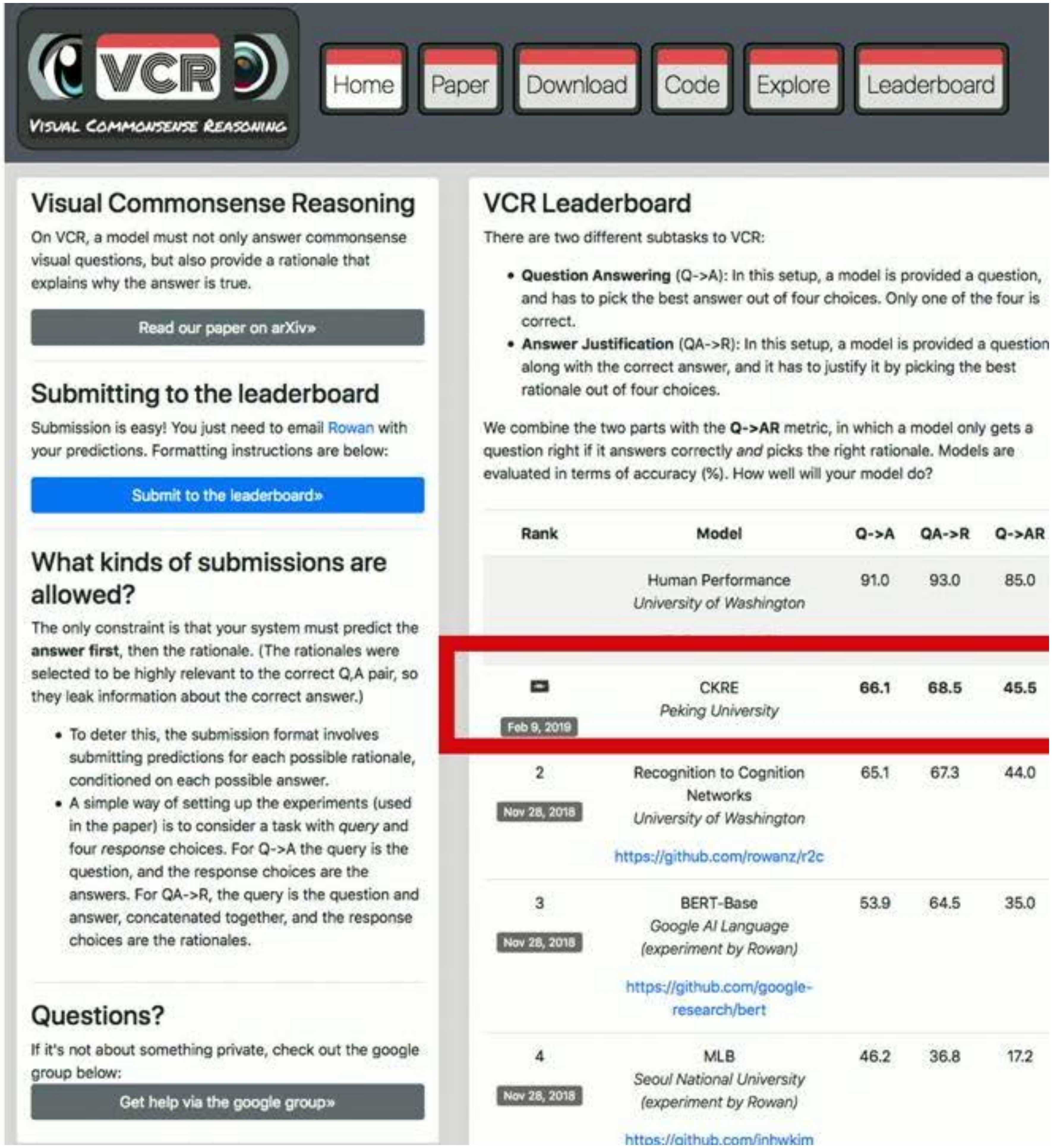
There are two different subtasks to VCR:

- **Question Answering** (Q->A): In this setup, a model is provided a question, and has to pick the best answer out of four choices. Only one of the four is correct.
- **Answer Justification** (QA->R): In this setup, a model is provided a question along with the correct answer, and it has to justify it by picking the best rationale out of four choices.

We combine the two parts with the **Q->AR** metric, in which a model only gets a question right if it answers correctly *and* picks the right rationale. Models are evaluated in terms of accuracy (%). How well will your model do?

Rank	Model	Q->A	QA->R	Q->AR
	Human Performance *University of Washington* (Zellers et al. '18)	91.0	93.0	85.0
Feb 9, 2019	CKRE *Peking University*	**66.1**	**68.5**	**45.5**
2 Nov 28, 2018	Recognition to Cognition Networks *University of Washington* https://github.com/rowanz/r2c	65.1	67.3	44.0
3 Nov 28, 2018	BERT-Base *Google AI Language (experiment by Rowan)* https://github.com/google-research/bert	53.9	64.5	35.0
4 Nov 28, 2018	MLB *Seoul National University (experiment by Rowan)* https://github.com/jnhwkim	46.2	36.8	17.2

visualcommonsense.com



VCR
Visual Commonsense Reasoning

Home | Paper | Download | Code | Explore | Leaderboard

Visual Commonsense Reasoning

On VCR, a model must not only answer commonsense visual questions, but also provide a rationale that explains why the answer is true.

Read our paper on arXiv»

Submitting to the leaderboard

Submission is easy! You just need to email Rowan with your predictions. Formatting instructions are below:

Submit to the leaderboard»

What kinds of submissions are allowed?

The only constraint is that your system must predict the **answer first**, then the rationale. (The rationales were selected to be highly relevant to the correct Q,A pair, so they leak information about the correct answer.)

- To deter this, the submission format involves submitting predictions for each possible rationale, conditioned on each possible answer.
- A simple way of setting up the experiments (used in the paper) is to consider a task with *query* and four *response* choices. For Q->A the query is the question, and the response choices are the answers. For QA->R, the query is the question and answer, concatenated together, and the response choices are the rationales.

Questions?

If it's not about something private, check out the google group below:

Get help via the google group»

VCR Leaderboard

There are two different subtasks to VCR:

- **Question Answering** (Q->A): In this setup, a model is provided a question, and has to pick the best answer out of four choices. Only one of the four is correct.
- **Answer Justification** (QA->R): In this setup, a model is provided a question along with the correct answer, and it has to justify it by picking the best rationale out of four choices.

We combine the two parts with the **Q->AR** metric, in which a model only gets a question right if it answers correctly *and* picks the right rationale. Models are evaluated in terms of accuracy (%). How well will your model do?

Rank	Model	Q->A	QA->R	Q->AR
	Human Performance *University of Washington*	91.0	93.0	85.0
Feb 9, 2019	CKRE *Peking University*	**66.1**	**68.5**	**45.5**
2 Nov 28, 2018	Recognition to Cognition Networks *University of Washington* https://github.com/rowanz/r2c	65.1	67.3	44.0
3 Nov 28, 2018	BERT-Base *Google AI Language (experiment by Rowan)* https://github.com/google-research/bert	53.9	64.5	35.0
4 Nov 28, 2018	MLB *Seoul National University (experiment by Rowan)* https://github.com/jnhwkim	46.2	36.8	17.2

What’s next?

Future work

- VCR is a new testbed for commonsense visual reasoning!
- What kinds of new models will do well?


VCR
Visual Commonsense Reasoning
VCR
Visual Commonsense Reasoning

Summary

SWAG

VCR

Visual Commonsense Reasoning

Summary

SWAG

VCR

Visual Commonsense Reasoning

R2C

UWNLP

PAUL G. ALLEN SCHOOL
OF COMPUTER SCIENCE & ENGINEERING

UNIVERSITY of WASHINGTON

Thanks all!!

More at rowanzellers.com

twitter: @rown

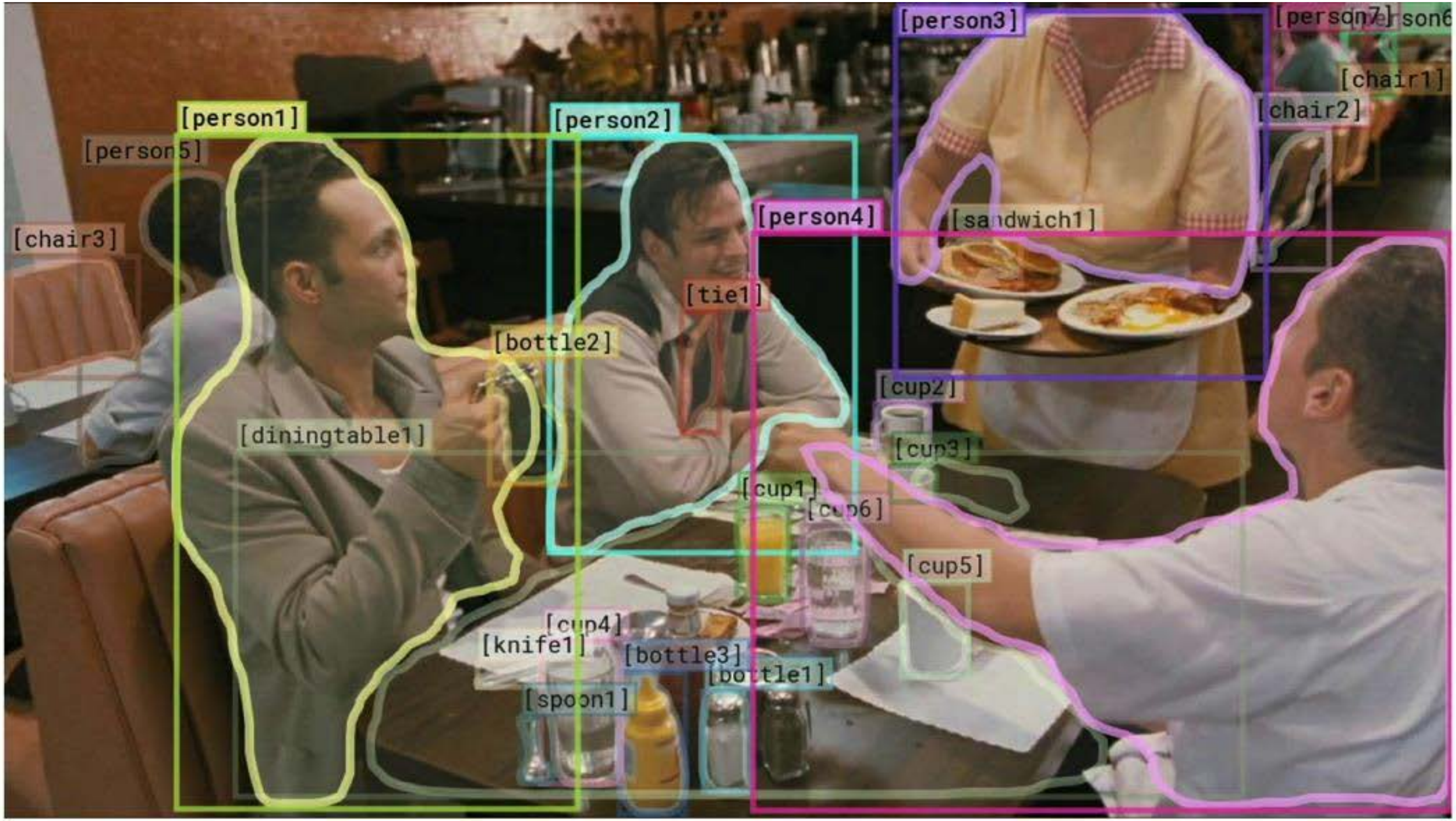
Me

Yonatan Bisk

Roy Schwartz

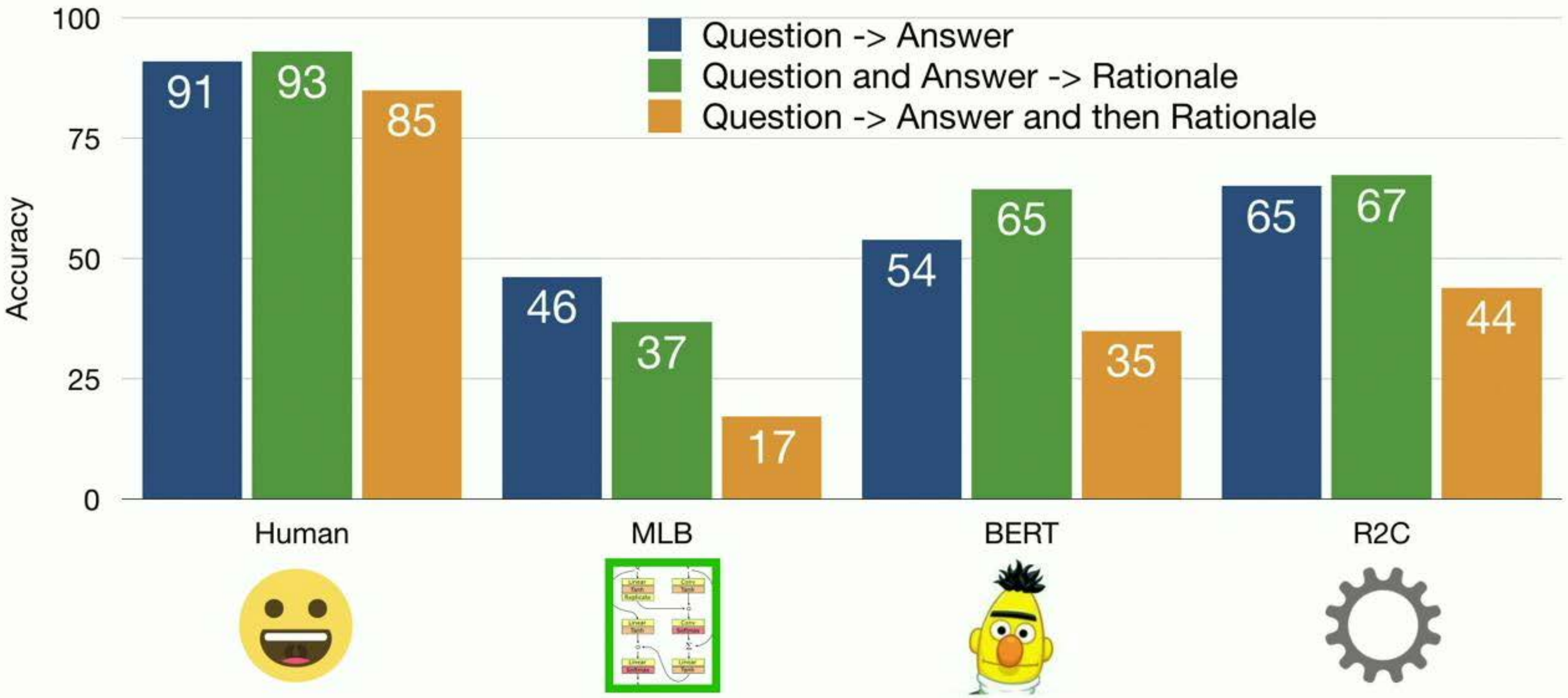
Ali Farhadi

Yejin Choi



Why is [person4 ] pointing at [person1 ]?

VCR Results



100
75
50
25
0
Accuracy
Question -> Answer
Question and Answer -> Rationale
Question -> Answer and then Rationale
91
93
85
46
37
17
54
65
35
65
67
44
Human
MLB
BERT
R2C